ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Εξόρυξη γνώσης για εξαγωγή πληροφορίας από τον παγκόσμιο ιστό με χρήση τεχνικών ψηφοφορίας και συσσωρευμένης γενίκευσης

Γεώργιος Βασιλείου Συγλέτος

ΑΘΗΝΑ

ΝΟΕΜΒΡΙΟΣ 2005

Εξόρυξη γνώσης για εξαγωγή πληροφορίας από τον παγκόσμιο ιστό με χρήση τεχνικών ψηφοφορίας και συσσωρευμένης γενίκευσης

Γεώργιος Βασιλείου Συγλέτος

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:
Μιχάλης Χατζόπουλος, Καθηγητής ΕΚΠΑ

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ ΠΑΡΑΚΟΛΟΥΘΗΣΗΣ:
Μιχάλης Χατζόπουλος, Καθηγητής ΕΚΠΑ
Παναγιώτης Σταματόπουλος, Επίκουρος Καθηγητής ΕΚΠΑ
Κωνσταντίνος Σπυρόπουλος, Διευθυντής Έρευνας, ΕΚΕΦΕ Δημόκριτος

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Μιχάλης Χατζόπουλος,
Καθηγητής ΕΚΠΑ

Κωνσταντίνος Χαλάτσης,
Καθηγητής ΕΚΠΑ

Παναγιώτης Σταματόπουλος,
Επίκουρος Καθηγητής ΕΚΠΑ

Μανόλης Κουμπαράκης,
Αναπληρωτής Καθηγητής ΕΚΠΑ

Ιωάννης Βλαχάβας,
Καθηγητής ΑΠΘ

Τίμος Σελλής,
Καθηγητής ΕΜΠ

Γεώργιος Βούρος,
Αναπληρωτής Καθηγητής
Πανεπιστημίου Αιγαίου

Ημερομηνία εξέτασης 04/11/2005

ΠΕΡΙΛΗΨΗ

Η ραγδαία εξάπλωση του Παγκοσμίου Ιστού και των άλλων υπηρεσιών του διαδικτύου τα τελευταία χρόνια επιτείνουν την ανάγκη ανάπτυξης συστημάτων που να βοηθούν τους χρήστες να εκμεταλλευτούν τον τεράστιο όγκο κειμένου που είναι διαθέσιμος στο διαδίκτυο. Τα συστήματα *εξαγωγής πληροφορίας (information extraction)*, δηλαδή συστήματα που εντοπίζουν και εξάγουν σχετικά τμήματα κειμένου από συλλογές κειμένων που αναφέρονται σε μια συγκεκριμένη θεματική περιοχή, δείχνουν ως μια ελπιδοφόρα διέξοδος για την αντιμετώπιση της πληροφοριακής αυτής έκρηξης. Η χρήση τεχνικών *μηχανικής μάθησης (machine learning)* διευκολύνει την ανάπτυξη συστημάτων εξαγωγής πληροφορίας, καθώς και μεταφερσιμότητά τους σε νέες θεματικές περιοχές ενδιαφέροντος. Η εξαγωγή πληροφορίας με χρήση τεχνικών μηχανικής μάθησης είναι ένα τυπικό πρόβλημα *εξόρυξης γνώσης από δεδομένα του παγκοσμίου ιστού (web mining)*, μιας και ο στόχος είναι η εκμάθηση κανόνων οι οποίοι θα εντοπίζουν αποτελεσματικά και θα εξάγουν σχετικά τμήματα κειμένου από έγγραφα του ιστού.

Σε αυτή τη διατριβή αναδεικνύεται η αποτελεσματικότητα του συνδυασμού συστημάτων εξαγωγής πληροφορίας με χρήση τεχνικών *ψηφοφορίας (voting)* και *συσσωρευμένης γενίκευσης (stacked generalization)* ή αλλιώς *συσσώρευσης (stacking)*. Το κίνητρο για το συνδυασμό πηγάζει από τη δυνατότητα για επίτευξη καλύτερης απόδοσης σε μετα-επίπεδο, εκμεταλλευόμενοι τη διαφορετικότητα στις προβλέψεις συστημάτων εξαγωγής πληροφορίας που χρησιμοποιούνται σε βασικό επίπεδο. Οι υπάρχουσες τεχνικές συνδυασμού επικεντρώνονται σε κοινά προβλήματα *ταξινόμησης (classification)*. Όμως η εξαγωγή πληροφορίας δεν είναι από τη φύση της ένα πρόβλημα ταξινόμησης. Προτείνεται λοιπόν μια νέα μεθοδολογία για το συνδυασμό συστημάτων εξαγωγής πληροφορίας μέσω ψηφοφορίας και συσσώρευσης. Η προτεινόμενη μεθοδολογία διευκολύνει το συνδυασμό πληθώρας συστημάτων εξαγωγής πληροφορίας αφού μόνο η έξοδός τους συνδυάζεται, αγνοώντας τις λεπτομέρειες υλοποίησης του κάθε συστήματος αλλά και το πώς κάθε σύστημα μοντελοποιεί το πρόβλημα της εξαγωγής πληροφορίας. Η εξαγωγή πληροφορίας μετατρέπεται τελικά σε ένα κοινό πρόβλημα ταξινόμησης σε μετα-επίπεδο, επιτρέποντας την εφαρμογή τεχνικών ψηφοφορίας και συσσωρευμένης γενίκευσης.

Αρχικά διερευνάται η αποτελεσματικότητα της *ψηφοφορίας* για το συνδυασμό συστημάτων εξαγωγής πληροφορίας. Εκτενή πειράματα πραγματοποιήθηκαν σε ποικιλία θεματικών περιοχών χρησιμοποιώντας γνωστά συστήματα εξαγωγής πληροφορίας σε βασικό επίπεδο. Τα αποτελέσματα αναδεικνύουν την αποτελεσματικότητα της ψηφοφορίας με χρήση πιθανοτήτων ορθότητας στις προβλέψεις των συστημάτων του βασικού επιπέδου, εφόσον τεθεί ένα όριο στην πιθανότητα ορθότητας για την αποδοχή ή όχι μιας πρόβλεψης σε μετα-επίπεδο. Στις περισσότερες θεματικές περιοχές, η ψηφοφορία ξεπέρασε σε απόδοση τα καλύτερα συστήματα εξαγωγής πληροφορίας του βασικού επιπέδου.

Κατόπιν διερευνάται η αποτελεσματικότητα της *συσσώρευσης* για το συνδυασμό συστημάτων εξαγωγής πληροφορίας. Η βασική ιδέα είναι να συνδυάσουμε πολλαπλά συστήματα εξαγωγής πληροφορίας με έναν κοινό ταξινομητή σε μετα-επίπεδο, όπως είναι ένας ταξινομητής *δέντρων απόφασης (decision trees)* ή ένας Naïve-Bayes ταξινομητής. Διασταυρωμένη επικύρωση λαμβάνει χώρα στο σύνολο δεδομένων του βασικού επιπέδου, το οποίο αποτελείται από επισημειωμένα κείμενα, για τη δημιουργία ενός συνόλου δεδομένων σε μετα-επίπεδο το οποίο αποτελείται από διανύσματα χαρακτηριστικών. Ένας κοινός ταξινομητής εκπαιδεύεται στη συνέχεια χρησιμοποιώντας τα νέα διανύσματα. Τα αποτελέσματα αναδεικνύουν την αποτελεσματικότητα της συσσώρευσης με χρήση πιθανοτήτων ορθότητας στις προβλέψεις των συστημάτων του βασικού επιπέδου. Η αποτελεσματικότητα της συσσώρευσης είναι καθολική σε όλες τις θεματικές περιοχές, ξεπερνώντας πάντα σε απόδοση τα καλύτερα συστήματα του βασικού επιπέδου. Σε σχέση

με την ψηφοφορία, η συσσώρευση επιτυγχάνει συγκρίσιμη ή και καλύτερη απόδοση και σε κάθε περίπτωση επιτυγχάνει προβλέψεις με μεγαλύτερη ακρίβεια σε μετα-επίπεδο.

Ιδιαίτερη έμφαση δόθηκε επίσης στην ανάλυση των αποτελεσμάτων που επιτυγχάνονται από την ψηφοφορία και τη συσσώρευση, με στόχο να διερευνηθούν οι πτυχές της επιτυχίας των τεχνικών αυτών για προβλήματα εξαγωγής πληροφορίας. Η ανάλυση έδειξε ότι οι τεχνικές αυτές εκμεταλλεύονται επιτυχώς τη διαφορετικότητα στις προβλέψεις των συστημάτων του βασικού επιπέδου, οδηγώντας σε καλύτερη απόδοση σε μετα-επίπεδο. Η συσσώρευση αποδεικνύεται αποτελεσματικότερη από την ψηφοφορία, ακόμα στην περίπτωση ταύτισης των προβλέψεων όλων των συστημάτων του βασικού επιπέδου.

Αισιοδοξώ ότι η διατριβή αυτή συνεισφέρει στην αναγνώριση της μεγάλης δυναμικής των τεχνικών συνδυασμού, προς την κατεύθυνση του εντοπισμού με ακρίβεια σχετικής πληροφορίας στον τεράστιο όγκο κειμένου ο οποίος βρίσκεται σε ψηφιακή μορφή, προσδοκώντας σε μια μέθοδο εύκολα προσαρμόσιμη σε νέες θεματικές περιοχές.

ΘΕΜΑΤΙΚΕΣ ΠΕΡΙΟΧΕΣ: εξαγωγή πληροφορίας, μηχανική μάθηση

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: εξαγωγή, μάθηση, ψηφοφορία, συσσωρευμένη γενίκευση, ιστός

ABSTRACT

The proliferation of the World Wide Web and the other Internet services in the past few years intensifies the need for developing systems that help users to cope with the enormous amount of text that is available online. *Information extraction* systems, that is, systems that locate and pull relevant fragments out of domain-specific collections of text documents, seem to be a promising way to deal with the information explosion. *Machine learning* techniques facilitate the development of information extraction systems and their portability to new domains of interest. Information extraction using machine learning techniques is a typical *Web mining* problem, since the task is to learn extraction rules that can effectively recognize relevant text fragments within Web documents.

This dissertation demonstrates the effectiveness of combining information extraction systems using *voting* and *stacked generalization (stacking)*. The motivation derives from the opportunity to obtain higher extraction performance at meta-level, by exploiting the disagreement in the predictions of the information extraction systems that are employed at base-level. Existing combination techniques primarily focus on classification. However, information extraction is not naturally a classification problem. A new methodology is proposed for combining information extraction systems through voting and stacking. The proposed methodology facilitates the combination of a wide range of systems, since only their output is combined, without taking into account how each system is implemented or models the extraction task. Information extraction is transformed to a common classification problem at meta-level, allowing the applicability of voting and stacking techniques.

The effectiveness of *voting* is initially investigated for combining multiple information extraction systems at meta-level. Extensive experiments were performed in a variety of domains using well known information extraction systems at base-level. The results demonstrate the effectiveness of voting with probabilistic estimates of correctness in the output of the base-level systems, as long as a probability threshold is set for deciding whether to accept a prediction at meta-level. Voting was effective on most domains in the experiments, outperforming the best base-level systems.

The effectiveness of *stacking* is then investigated for combining multiple information extraction systems at meta-level. The basic idea is to combine well known information extraction systems with a common classifier at meta-level, such as a decision-tree classifier or a Naïve Bayes classifier. Cross-validation takes place on the base-level dataset, which consists of text documents annotated with relevant information, in order to create a meta-level dataset that consists of feature vectors. A common classifier is then trained using the new vectors. Results demonstrate the effectiveness of stacking using probabilities in the output of the base-level systems. Stacking was consistently effective in all examined domains, always outperforming the best base-level information extraction systems. Comparing against voting, stacking performs comparably or better, while always obtaining more accurate predictions at meta-level.

Particular emphasis was also given to analyzing the results obtained by voting and stacking, aiming to investigate the sources of their success in information extraction tasks. The analysis showed that voting and stacking successfully exploit the disagreement in the output of the base-level systems, towards better results at meta-level. Stacking, however, proved to be better than voting, even when all base-level systems predict identically.

This dissertation contributes to the direction of realizing the high potential of combination methods in the context of accurately identifying relevant items of information on the abundant of computerized text, aiming at a method easily adapted to new domains.

SUBJECT AREAS: information extraction, machine learning

KEYWORDS: extraction, learning, voting, stacked generalization, web

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διατριβή πραγματοποιήθηκε με τη βοήθεια υποτροφίας που χορηγήθηκε από το Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών του Εθνικού Κέντρου Έρευνας Φυσικών Επιστημών (Ε.Κ.Ε.Φ.Ε.) “Δημόκριτος”. Ευχαριστώ όλους όσους συντέλεσαν στην επιτυχή ολοκλήρωση της διατριβής αυτής.

Ευχαριστώ ιδιαίτερα τον κ. Κωνσταντίνο Σπυρόπουλο, Διευθυντή Έρευνας του Ε.Κ.Ε.Φ.Ε. “Δημόκριτος” για τα χρήσιμα σχόλιά του καθ’ όλη τη διάρκεια του διδακτορικού, στο τελικό κείμενο της διατριβής, καθώς και για το γεγονός ότι εξασφάλισε χρηματοδότηση για μια σειρά από συνέδρια τα οποία ανέδειξαν την ερευνητική μου εργασία σε διεθνή επίπεδο. Ευχαριστώ επίσης τον κ. Γεώργιο Παλιούρα, Ερευνητή Γ’, για την καθοδήγησή του καθ’ όλη τη διάρκεια του διδακτορικού και τη συνεισφορά του στη βελτίωση του τρόπου συγγραφής των επιστημονικών μου άρθρων. Ευχαριστώ και τον Βαγγέλη Καρκαλέτση, Ερευνητή Β, για την καθοδήγησή του σε θέματα επεξεργασίας κειμένου και εξαγωγής πληροφορίας στην αρχή του διδακτορικού, καθώς και για την παραχώρηση μιας συλλογής κειμένων από τις πέντε που χρησιμοποίησα για τη διεξαγωγή πειραμάτων. Ευχαριστώ επίσης τα υπόλοιπα δύο μέλη της τριμελούς μου επιτροπής από το Πανεπιστήμιο Αθηνών, τον καθηγητή κ. Μιχάλη Χατζόπουλο και τον επίκουρο καθηγητή κ. Παναγιώτη Σταματόπουλο για τη δημιουργική συνεργασία που είχαμε.

Ευχαριστώ τους Γεώργιο Πετάση, Γεώργιο Σάκη, Σέργιο Πετρίδη και Αλέξανδρο Βαλαράκο, για τη βοήθειά τους σε διάφορα τεχνικά θέματα. Ευχαριστώ επίσης τον Dr. Fabio Ciravegna, καθηγητή στο Πανεπιστήμιο του Sheffield στην Αγγλία, για την ευγενική παραχώρηση του συστήματος εξαγωγής πληροφορίας από κείμενα, $(LP)^2$, που χρησιμοποίησα για τη διεξαγωγή πειραμάτων στα πλαίσια της διατριβής αυτής. Ευχαριστώ επίσης τον Dr. Dayne Freitag, απόφοιτο του Πανεπιστημίου Carnegie Mellon των Η.Π.Α. για την ευγενική παραχώρηση των τριών από τις συνολικά πέντε συλλογές κειμένων που χρησιμοποίησα για πειράματα στα πλαίσια της διατριβής αυτής.

Τέλος, ένα πολύ μεγάλο ευχαριστώ στους γονείς μου, που με στήριξαν -και οικονομικά- από το πρώτο έτος των βασικών σπουδών μου έως και σήμερα, συνεισφέροντας σημαντικά στην επιτυχή ολοκλήρωση της διατριβής αυτής.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

Η παρούσα διατριβή μελετά την αποτελεσματικότητα τεχνικών *ψηφοφορίας* και *συσσωρευμένης γενίκευσης* για το συνδυασμό συστημάτων εξαγωγής πληροφορίας. Στόχος είναι η επίτευξη καλύτερης απόδοσης στην εξαγωγή πληροφορίας σε μετα-επίπεδο, σε σχέση με ένα σύνολο συστημάτων που είναι διαθέσιμα σε βασικό επίπεδο. Προτείνεται και αξιολογείται μια νέα μεθοδολογία συνδυασμού συστημάτων η οποία δεν περιορίζεται σε κοινά προβλήματα ταξινόμησης όπως οι υπάρχουσες τεχνικές συνδυασμού. Από την άλλη πλευρά, ο συνδυασμός συστημάτων εξαγωγής πληροφορίας έχει βρει αμελητέα ανταπόκριση μέχρι τώρα στην ερευνητική κοινότητα.

Η Ενότητα 1.1 πραγματοποιεί μια σύντομη επισκόπηση στο αντικείμενο της παρούσας διατριβής το οποίο συνδυάζει τις ερευνητικές περιοχές της εξαγωγής πληροφορίας και της μηχανικής μάθησης. Η Ενότητα 1.2 περιγράφει τη συνεισφορά της διατριβής ενώ η Ενότητα 1.3 αναφέρει τις δημοσιεύσεις που επιτευχθήκαν κατά τη διάρκεια εκπόνησής της. Τέλος, η Ενότητα 1.4 περιγράφει τη διάρθρωση του υπολοίπου της διατριβής.

1.1 Σύντομη επισκόπηση στο αντικείμενο

Η ραγδαία εξάπλωση του ιστού τα τελευταία χρόνια συνίσταται στην αλματώδη αύξηση των χρηστών του διαδικτύου καθώς και των εγγράφων κειμένου που δημοσιεύονται καθημερινά στις σελίδες των χρηστών του. Η αναζήτηση της επιθυμητής πληροφορίας μέσα από τον τεράστιο όγκο σελίδων που είναι δημοσιευμένες στον ιστό, αποτελεί ένα σημαντικό πρόβλημα κατά την αλληλεπίδραση με το διαδίκτυο. Η αποτελεσματική διαχείριση του τεράστιου αυτού όγκου δεδομένων αποτελεί ιδανικό κίνητρο για την έρευνα τεχνικών εξόρυξης γνώσης από δεδομένα. Ο όρος *εξόρυξη γνώσης από δεδομένα (data mining)* αναφέρεται στην εφαρμογή τεχνικών *μηχανικής μάθησης (machine learning)* σε μεγάλο όγκο δεδομένων για την *ανακάλυψη γνώσης (knowledge discovery)* από τα δεδομένα αυτά. Η εξόρυξη γνώσης από δεδομένα του ιστού καλείται *Web Mining* [43] και αποτελεί σημείο συνάντησης ερευνητικών περιοχών όπως η μηχανική μάθηση, οι βάσεις δεδομένων, η επεξεργασία φυσικής γλώσσας, η ανάκτηση πληροφορίας, η εξαγωγή πληροφορίας και η μοντελοποίηση χρηστών.

Η *εξαγωγή πληροφορίας (information extraction)* είναι μια νέα ερευνητική περιοχή που αναπτύσσεται ταχύτατα τα τελευταία χρόνια και τοποθετείται μεταξύ των περιοχών της *ανάκτησης πληροφορίας (information retrieval)* και της *επεξεργασίας φυσικής γλώσσας*

(natural language processing). Αντίθετα με την ανάκτηση πληροφορίας όπου το πρόβλημα είναι ο εντοπισμός των εγγράφων εκείνων που περιέχουν την επιθυμητή πληροφορία, στην εξαγωγή πληροφορίας το πρόβλημα είναι να εντοπιστούν τα σχετικά τμήματα κειμένου μέσα σε ένα έγγραφο που ανήκει σε μια θεματική περιοχή. Αντίθετα επίσης με την επεξεργασία φυσικής γλώσσας που περιλαμβάνει την κατανόηση σε βάθος ενός εγγράφου, η εξαγωγή πληροφορίας θεωρείται ως ένα είδος περιορισμένης κατανόησης ενός εγγράφου, σε τόσο βάθος όσο απαιτείται για να εντοπιστεί η σχετική πληροφορία μέσα σε αυτό. Το ενδιαφέρον πηγάζει από το γεγονός ότι η πληροφορία που εξάγεται τοποθετείται σε ένα προκαθορισμένο *σχεδιότυπο (template)* ή μια βάση δεδομένων, διευκολύνοντας την περαιτέρω διαχείρισή της.

Η εξαγωγή πληροφορίας από έγγραφα κειμένου του παγκοσμίου ιστού πραγματοποιείται τυπικά μέσω ειδικών προγραμμάτων εξαγωγής τα οποία καλούνται *wrappers* και είναι κανόνες που αναγνωρίζουν με μεγάλη ακρίβεια σχετικά τμήματα κειμένου σε ένα έγγραφο του ιστού. Η χειρονακτική κατασκευή τέτοιων προγραμμάτων [21] είναι μια επίπονη και χρονοβόρα διαδικασία που απαιτεί μεγάλο βαθμό εμπειρογνωμοσύνης. Τεχνικές μηχανικής μάθησης έχουν χρησιμοποιηθεί για την εκμάθηση κανόνων *wrappers* [20, 29, 32, 58, 69, 82]. Κατά συνέπεια, η εξαγωγή πληροφορίας από έγγραφα του ιστού με χρήση μηχανικής μάθησης είναι ένα πρόβλημα εξόρυξης γνώσης από δεδομένα του ιστού, καθώς η χρήση μηχανικής μάθησης αποσκοπεί στην *εξόρυξη* κανόνων οι οποίοι θα μπορούν να αναγνωρίζουν και να εξάγουν σχετικά τμήματα σε ένα έγγραφο κειμένου του ιστού.

Ένα σημαντικό μειονέκτημα της χρήσης των ειδικών προγραμμάτων *wrappers* είναι η αδυναμία τους στην εξαγωγή πληροφορίας από σελίδες μη αυστηρά δομημένες. Στον ιστό είναι όμως αρκετά συχνό το φαινόμενο ύπαρξης σελίδων με χαλαρά δομημένο ή εντελώς αδόμητο (ελεύθερο) περιεχόμενο. Από την άλλη πλευρά, τα συστήματα εξαγωγής πληροφορίας που αναπτύχθηκαν στα πλαίσια των συνέδριων MUC (Message Understanding Conferences, [35, 36]), παρόλο που στοχεύουν στην εξαγωγή πληροφορίας από αδόμητο κείμενο, δεν έχουν βρει ανταπόκριση στο χώρο του ιστού. Αυτό οφείλεται αφενός στο υψηλό κόστος ανάπτυξης και προσαρμογής των συστημάτων αυτών σε νέες περιοχές, και αφετέρου στην αδυναμία εκμετάλλευσης εξτρα-γλωσσικής πληροφορία (HTML/XML) που είναι διαθέσιμη στις σελίδες του ιστού.

Πρόσφατη έρευνα παροτρύνει την ανάπτυξη *προσαρμοστικών (adaptive)* συστημάτων εξαγωγής πληροφορίας [24, 26], τα οποία θα είναι σε θέση να χειριστούν κείμενα διαφορετικού τύπου δόμησης, από υψηλά δομημένο σε σχεδόν ελεύθερο κείμενο. Τα

συστήματα που περιγράφονται στις εργασίες [16, 24, 49, 50] χρησιμοποιούν αλγορίθμους μάθησης για την εκμάθηση κανόνων εξαγωγής με εφαρμογή τόσο σε δομημένο όσο και σε αδόμητο κείμενο.

Σημαντικό κίνητρο για την ανάπτυξη προσαρμοστικών συστημάτων εξαγωγής πληροφορίας αποτελεί η εκπλήρωση των στόχων του *σημασιολογικού ιστού (semantic web)*. Η χρήση προτύπων σημασιολογικής επισημείωσης κειμένων, τα οποία πρότυπα βασίζονται στη γλώσσα XML, σε συνδυασμό με τη χρήση *οντολογιών (ontologies)*, αποσκοπούν στη δημιουργία *περιεχομένου κατανοητού από τις μηχανές (machine readable content)*. Δυστυχώς, η συντριπτική πλειοψηφία των εγγράφων του ιστού δεν είναι σημασιολογικά επισημειωμένες, ενώ το κόστος της χειρονακτικής επισημείωσής τους είναι απαγορευτικό. Η εξαγωγή πληροφορίας ως μια αυτόματη μέθοδος εντοπισμού σχετικής πληροφορίας μέσα σε κείμενο μπορεί να παίξει σημαντικό ρόλο στη σημασιολογική επισημείωση κειμένου του ιστού και τον εμπλουτισμό οντολογιών με νέα πληροφορία [25, 113]. Η χρήση προσαρμοστικών συστημάτων αποτελεί απαίτηση για το σκοπό αυτό, καθώς ο ιστός περιέχει τόσο δομημένο όσο και αδόμητο κείμενο.

Η παρούσα διατριβή ερευνά την αποτελεσματικότητα τεχνικών *ψηφοφορίας (voting)* και *συσσωρευμένης γενίκευσης (stacked generalization)* ή αλλιώς *συσσώρευσης (stacking)* για το συνδυασμό προσαρμοστικών συστημάτων εξαγωγής πληροφορίας. Στόχος είναι η επίτευξη καλύτερης απόδοσης στην εξαγωγή σε μετα-επίπεδο, σε σχέση με ένα σύνολο συστημάτων εξαγωγής πληροφορίας που είναι διαθέσιμα σε βασικό επίπεδο. Ο συνδυασμός συστημάτων εξαγωγής πληροφορίας, όμως, έχει βρει αμελητέα ανταπόκριση μέχρι σήμερα στην ερευνητική κοινότητα.

Οι τεχνικές ψηφοφορίας και συσσωρευμένης γενίκευσης, όπως έχουν οριστεί μέχρι τώρα, απευθύνονται σε προβλήματα κοινής ταξινόμησης. Σε ένα πρόβλημα ταξινόμησης, κάθε παράδειγμα εκπαίδευσης αντιστοιχεί σε ένα διάνυσμα $< x_1 ... x_n , y >$, όπου $x_1 ... x_n$ είναι ένα σύνολο τιμών χαρακτηριστικών, ή αλλιώς γνωρισμάτων, και y είναι μια τιμή κλάσης η οποία περιγράφει ένα συγκεκριμένο γεγονός για μια θεματική περιοχή. Η συσσωρευμένη γενίκευση [120] πραγματεύεται την εκπαίδευση ενός ταξινομητή σε μετα-επίπεδο από τις προβλέψεις ενός συνόλου ταξινομητών που χρησιμοποιούνται σε βασικό επίπεδο. Για την ταξινόμηση ενός νέου διανύσματος $< x_1 ... x_n >$, οι προβλέψεις των ταξινομητών του βασικού επιπέδου για την τιμή κλάσης y σχηματίζουν ένα νέο διάνυσμα χαρακτηριστικών του οποίου την τιμή κλάσης αναθέτει ο ταξινομητής του μετα-επιπέδου. Μια διαδικασία *διασταυρωμένης επικύρωσης (cross-validation)* λαμβάνει χώρα στο αρχικό σύνολο των διανυσμάτων

χαρακτηριστικών για τη δημιουργία ενός συνόλου διανυσμάτων χαρακτηριστικών σε μετα-επίπεδο και την εκπαίδευση του αντίστοιχου ταξινομητή. Αντίθετα, δε λαμβάνει χώρα μάθηση κατά την απλή ψηφοφορία στις προβλέψεις των ταξινομητών του βασικού επιπέδου.

Οι παρούσες τεχνικές ψηφοφορίας και συσσωρευμένης γενίκευσης δε μπορούν να εφαρμοστούν άμεσα στο πρόβλημα της εξαγωγής πληροφορίας, διότι αυτό δεν είναι από τη φύση του ένα πρόβλημα κοινής ταξινόμησης [69, 108]. Οι αλγόριθμοι μηχανικής μάθησης που είναι σχεδιασμένοι για εξαγωγή πληροφορίας, τυπικά μαθαίνουν κανόνες οι οποίοι αναγνωρίζουν σχετικά τμήματα κειμένου, δηλαδή σχετικές ακολουθίες *λεκτικών μονάδων (tokens)*, μέσα σε ένα έγγραφο. Αντίθετα, προβλήματα γλωσσικής ανάλυσης, όπως η *επισημείωση μερών του λόγου (part-of speech tagging)* και η *αποσαφήνιση εννοιών-λέξεων (word-sense disambiguation)* είναι κλασικά προβλήματα ταξινόμησης, όπου σε κάθε λεκτική μονάδα στο κείμενο ανατίθεται μια ετικέτα από ένα προκαθορισμένο σύνολο. Στην περίπτωση αυτή, ένα διάνυσμα χαρακτηριστικών αντιστοιχεί σε κάθε λεκτική μονάδα, ενώ η τιμή κλάσης y αντιστοιχεί στη σωστή ετικέτα.

Ένα πρόβλημα εξαγωγής πληροφορίας μπορεί να μετατραπεί έμμεσα σε ένα πρόβλημα ταξινόμησης, ως σχετικών ή μη, όλων σχεδόν των δυνατών τμημάτων κειμένου που μπορούν να απαριθμηθούν σε ένα έγγραφο [48]. Αυτός ο τρόπος μοντελοποίησης του προβλήματος της εξαγωγής πληροφορίας δεν είναι φυσικός και οδηγεί σε σημαντική αύξηση του αριθμού των υποψηφίων τμημάτων κειμένου προς ταξινόμηση, και της μεγάλης δυσαναλογίας αρνητικών και θετικών παραδειγμάτων.

Ο Πίνακας 1.1 δείχνει τα παραδείγματα εκπαίδευσης που κατασκευάζονται από ένα υποθετικό τμήμα μιας σελίδας που περιγράφει φορητούς ηλεκτρονικούς υπολογιστές. Ο Πίνακας δείχνει ότι κατασκευάζεται μόνο ένα θετικό παράδειγμα εκπαίδευσης και πληθώρα αρνητικών παραδειγμάτων.

Πίνακας 1.1 Παράδειγμα μοντελοποίησης της εξαγωγής πληροφορίας κατά Freitag [48]. Έστω ότι το “256 MB” είναι ένα σχετικό παράδειγμα του πεδίου *ram*.

Τμήμα κειμένου:	…processor <b> **256 MB** SDRAM…	
Θετικά παραδείγματα:	256 MB	
Αρνητικά παραδείγματα:	processor processor processor <b> 256 MB …	256 MB 256 MB SDRAM MB SDRAM …

Αντίθετα, συστήματα εξαγωγής πληροφορίας όπως αυτά που περιγράφονται στις εργασίες [24, 49, 82] μοντελοποιούν την εξαγωγή πληροφορίας ως ένα πρόβλημα *εντοπισμού ορίων (boundary detection)* μέσα στο κείμενο. Ένα *όριο* ορίζεται ως το εικονικό διάστημα μεταξύ δύο γειτονικών λεκτικών μονάδων. Στην περίπτωση αυτή τα αρχικά και τελικά όρια των σχετικών τμημάτων κειμένου σε ένα έγγραφο πρέπει να αναγνωριστούν και στη συνέχεια να εξαχθεί το περιεχόμενο ανάμεσα στα όρια αυτά. Στον Πίνακα 1.1, το όριο μεταξύ των “<b>” και “256” καθώς κι εκείνο μεταξύ “MB” και “SDRAM” είναι το αρχικό και τελικό όριο αντίστοιχα του τμήματος “256 MB”. Πλήθος άλλων προσεγγίσεων [16, 50, 106] μαθαίνουν κανόνες που εξάγουν ολόκληρα τμήματα κειμένου από ένα έγγραφο.

Επομένως παρουσιάζει ιδιαίτερο ενδιαφέρον η προσθήκη μιας νέας μεθοδολογίας συνδυασμού συστημάτων εξαγωγής πληροφορίας, η οποία θα ταιριάζει περισσότερο στη φύση του προβλήματος της εξαγωγής πληροφορίας και δεν θα περιορίζεται σε κοινά προβλήματα ταξινόμησης όπως οι υπάρχουσες τεχνικές συνδυασμού.

1.2 Συνεισφορά

Η παρούσα διατριβή προτείνει μια νέα μεθοδολογία για το συνδυασμό συστημάτων εξαγωγής πληροφορίας με χρήση τεχνικών *ψηφοφορίας* και *συσσωρευμένης γενίκευσης*, ανεξάρτητα από τον τρόπο μοντελοποίησης του προβλήματος της εξαγωγής πληροφορίας από κάθε σύστημα. Τα επισημειωμένα με την επιθυμητή πληροφορία έγγραφα που είναι διαθέσιμα σε βασικό επίπεδο μετασχηματίζονται κατάλληλα σε διανύσματα χαρακτηριστικών σε μετα-επίπεδο, επιτρέποντας έτσι την άμεση εφαρμογή τεχνικών ψηφοφορίας και συσσωρευμένης γενίκευσης.

Τεχνικές *ψηφοφορίας* ορίζονται και αξιολογούνται για το συνδυασμό διαφορετικών συστημάτων εξαγωγής πληροφορίας. Οι τεχνικές αυτές βασίζονται τόσο στη χρήση ονομαστικών τιμών στις προβλέψεις των συστημάτων εξαγωγής πληροφορίας του βασικού επιπέδου *(πλειοψηφική ψηφοφορία)*, όσο και στη χρήση πιθανοτήτων ορθότητας *(πιθανοτική ψηφοφορία)*. Η αξιολόγηση αναδεικνύει την αποτελεσματικότητα της πιθανοτικής ψηφοφορίας, εφόσον τεθεί ένα κατώφλι στην πιθανότητα για την αποδοχή ή όχι μιας πρόβλεψης σε μετα-επίπεδο. Στις περισσότερες θεματικές περιοχές ενδιαφέροντος, η ψηφοφορία οδήγησε σε καλύτερα αποτελέσματα στην εξαγωγή πληροφορίας από τα συστήματα του βασικού επιπέδου.

Τεχνικές *συσσωρευμένης γενίκευσης* ορίζονται και αξιολογούνται για το συνδυασμό διαφορετικών συστημάτων εξαγωγής πληροφορίας. Η βασική ιδέα είναι να

συνδυάσουμε συστήματα εξαγωγής πληροφορίας με έναν κοινό ταξινομητή σε μετα-επίπεδο, όπως είναι ένας ταξινομητής δέντρων απόφασης ή ένας Naïve-Bayes ταξινομητής. Οι τεχνικές συσσωρευμένης γενίκευσης βασίζονται επίσης στη χρήση ονομαστικών τιμών *(συσσωρευμένη γενίκευση με ονομαστικές τιμές)* στις προβλέψεις των συστημάτων του βασικού επιπέδου, όσο και στη χρήση πιθανοτήτων ορθότητας *(συσσωρευμένη γενίκευση με πιθανότητες)*. Η αξιολόγηση αναδεικνύει την καθολική αποτελεσματικότητα της συσσωρευμένης γενίκευσης με πιθανότητες σε όλες τις θεματικές περιοχές, οδηγώντας σε καλύτερα αποτελέσματα εξαγωγής από τα συστήματα του βασικού επιπέδου. Η συσσωρευμένη γενίκευση οδηγεί σε συγκρίσιμα ή και καλύτερα αποτελέσματα σε σχέση με την ψηφοφορία, και σε κάθε περίπτωση οδηγεί σε πιο ακριβείς προβλέψεις στην εξαγωγή πληροφορίας σε μετα-επίπεδο.

Τέλος, η διατριβή αυτή προσφέρει μια ανάλυση σε βάθος των αποτελεσμάτων που επιτυγχάνονται από τις τεχνικές ψηφοφορίας και συσσωρευμένης γενίκευσης, με στόχο τη διερεύνηση των πτυχών της επιτυχίας των τεχνικών αυτών. Παρόλο που όλες οι τεχνικές συνδυασμού εκμεταλλεύονται επιτυχώς τη διαφορετικότητα στις προβλέψεις των συστημάτων του βασικού επιπέδου, η συσσωρευμένη γενίκευση, με χρήση πιθανοτήτων, αποδεικνύεται ελαφρώς καλύτερη ακόμα και στην περίπτωση που όλα τα συστήματα ταυτίζονται στις προβλέψεις τους. Η ανάλυση έδειξε επίσης ότι συστήματα εξαγωγής πληροφορίας με μεγαλύτερη *ανάκληση* στις προβλέψεις τους θα πρέπει γενικότερα να προτιμώνται κατά το συνδυασμό, καθώς η *ακρίβεια* βελτιώνεται πάντα από τη συσσωρευμένη γενίκευση σε μετα-επίπεδο.

Η παρούσα διατριβή συνεισφέρει στην κατεύθυνση της εκπλήρωσης του στόχου του σημασιολογικού ιστού για τη δημιουργία περιεχομένου κατανοήσιμου από τις μηχανές των υπολογιστών, καθώς αναδεικνύει την αποτελεσματικότητα του συνδυασμού διαφορετικών προσαρμοστικών συστημάτων εξαγωγής, τα οποία έχουν εφαρμογή τόσο σε ισχυρά δομημένο όσο και σε λιγότερο δομημένο κείμενο του ιστού.

Παρόλο που το αρχικό ενδιαφέρον ήταν για δεδομένα του παγκοσμίου ιστού, η ιδέα του συνδυασμού διαφορετικών συστημάτων εξαγωγής πληροφορίας είναι αρκετά γενική και μπορεί να εφαρμοστεί και πέραν του ιστού. Για το λόγο αυτό χρησιμοποιήθηκαν κατά την αξιολόγηση και συλλογές κειμένων που δεν προέρχονται από το χώρο του παγκοσμίου ιστού, αλλά και από τον ευρύτερο χώρο του διαδικτύου όπως οι *ομάδες συζητήσεων (newsgroups)*. Η αξιολόγηση αναδεικνύει την αποτελεσματικότητα των τεχνικών ψηφοφορίας και συσσωρευμένης γενίκευσης και στην περίπτωση αυτή.

Η ευρύτερη συνεισφορά της διατριβής αυτής είναι η κατανόηση της μεγάλης δυναμικής των μεθόδων συνδυασμού συστημάτων εντοπισμού σχετικής πληροφορίας μέσα από κείμενα, με δυνατότητες μελλοντικής εφαρμογής και σε χώρους εκτός του παγκοσμίου ιστού και του διαδικτύου γενικότερα.

1.3 Δημοσιεύσεις

Στα πλαίσια της διατριβής προέκυψαν οι παρακάτω δημοσιεύσεις:

Sigletos, G., Paliouras, G., Spyropoulos, C.D., Hatzopoulos, M., Combining information extraction systems using voting and stacked generalization, *Journal of Machine Learning Research*, MIT Press/Microtome publishing, 6, 1751-1782, 2005.

(ISI 2004 Impact factor: 5.952 – το μεγαλύτερο από όλα τα περιοδικά της Τεχνητής Νοημοσύνης και το δεύτερο μεγαλύτερο από όλα τα περιοδικά της Επιστήμης Υπολογιστών)

Sigletos, G., Voting and stacking for information extraction: Extended results, *Technical Report DEMO 2005/3*, NCSR Demokritos, 2005 (συμπληρωματικό του προηγούμενου).

Sigletos, G., Paliouras, G., Spyropoulos, C.D., Stamatopoulos, T., Stacked generalization for information extraction, *In Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*, Valencia, Spain, IOS Press, 549-553, 2004.

Sigletos, G., Paliouras, G., Spyropoulos, C.D., Hatzopoulos, M., Mining Web sites using wrapper induction, named entities and post-processing, *Web Mining: From Web to Semantic Web*, LNAI 3209, Springer, 97-112, 2004.

Sigletos, G., Paliouras, G., Spyropoulos, C.D., Stamatopoulos, T., Meta-learning beyond classification: a framework for information extraction from the Web, *Workshop on Adaptive Text Extraction and Mining, held in conjunction with the 14th European Conference on Machine Learning (ECML)*, Dubrovnik, Croatia, 2003.

Sigletos, G., et al., Annotating Web pages for the needs of Web information extraction applications. Short paper in the Proceedings of the 12th International World Wide Web (WWW) Conference, Budapest, Hungary, 2003.

Sigletos, G., Paliouras, G., Karkaletsis, V., Role identification from free text using hidden Markov models, *Methods and applications of Artificial Intelligence*, LNAI 2308, Springer, 167-178, 2002.

1.4 Διάρθρωση της διατριβής

Το υπόλοιπο της διατριβής είναι διαρθρωμένο ως εξής:

Το Κεφάλαιο 2 περιγράφει τις βασικές έννοιες γύρω από τις περιοχές της μηχανικής μάθησης, του συνδυασμού αλγορίθμων σε μετα-επίπεδο και της εξαγωγής πληροφορίας, ενώ γίνεται εκτενής επισκόπηση στη σχετική βιβλιογραφία των περιοχών αυτών. Εκτενής επισκόπηση στη βιβλιογραφία πραγματοποιείται και στο χώρο του συνδυασμού διαφορετικών συστημάτων σε μετα-επίπεδο για προβλήματα εξαγωγής πληροφορίας, που πραγματεύεται η διατριβή αυτή.

Το Κεφάλαιο 3 περιγράφει αρχικά τις θεματικές περιοχές ενδιαφέροντος, τους αλγορίθμους που αξιολογήθηκαν σε βασικό και σε μετα-επίπεδο και τη μεθοδολογία αξιολόγησης. Επίσης, παρουσιάζονται και αναλύονται τα αποτελέσματα αξιολόγησης των συστημάτων εξαγωγής πληροφορίας του βασικού επιπέδου, ενώ διερευνάται εάν υπάρχουν περιθώρια βελτίωσης των αποτελεσμάτων αυτών σε μετα-επίπεδο.

Τα Κεφάλαια 4 και 5 ορίζουν και αξιολογούν τεχνικές *ψηφοφορίας* και *συσσωρευμένης γενίκευσης*, αντίστοιχα, για το συνδυασμό διαφορετικών συστημάτων εξαγωγής πληροφορίας. Οι τεχνικές αυτές βασίζονται τόσο στη χρήση ονομαστικών τιμών στις προβλέψεις των συστημάτων εξαγωγής πληροφορίας του βασικού επιπέδου, όσο και στη χρήση πιθανοτήτων ορθότητας στις προβλέψεις αυτές.

Το Κεφάλαιο 6 συγκρίνει τις τεχνικές ψηφοφορίας και συσσωρευμένης γενίκευσης για το συνδυασμό διαφορετικών συστημάτων εξαγωγής πληροφορίας, όπως περιγράφτηκαν στα προηγούμενα δύο κεφάλαια. Ταυτόχρονα πραγματοποιείται ανάλυση σε βάθος των αποτελεσμάτων που επιτυγχάνουν οι τεχνικές αυτές.

Τέλος, το Κεφάλαιο 7 συνοψίζει τα συμπεράσματα της παρούσας διατριβής, ενώ συζητούνται βελτιώσεις και επεκτάσεις της προτεινόμενης μεθοδολογίας.

ΚΕΦΑΛΑΙΟ 2

ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΚΑΙ ΕΠΙΣΚΟΠΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

Η τεχνητή νοημοσύνη και ειδικότερα η *μηχανική μάθηση* έχει αποκτήσει μεγάλο εύρος εφαρμογής στον παγκόσμιο ιστό τα τελευταία χρόνια. Κύριος στόχος είναι η αντιμετώπιση του προβλήματος της *υπερ-πληροφόρησης*, μέσω της ανάπτυξης συστημάτων τα οποία θα μπορούν αυτόματα να φιλτράρουν τον ολοένα και αυξανόμενο όγκο δεδομένων του ιστού, αναζητώντας σχετική πληροφορία για τον τελικό χρήστη. Η χρήση μηχανικής μάθησης στην ανάπτυξη συστημάτων *εξαγωγής πληροφορίας* που θα εντοπίζουν αυτόματα και θα εξάγουν σχετική πληροφορία από σελίδες του ιστού, συνιστά μια πολύ καλή προοπτική για την αντιμετώπιση της υπερ-πληροφόρησης.

Η Ενότητα 2.1 περιγράφει κάποιες βασικές έννοιες μηχανικής μάθησης. Οι Ενότητες 2.2 και 2.3 περιγράφουν τις μεθοδολογίες συνδυασμού πολλαπλών ταξινομητών μέσω ψηφοφορίας και συσσωρευμένης γενίκευσης αντίστοιχα, ενώ γίνεται επισκόπηση στη διεθνή βιβλιογραφία όπου αναφέρονται και εναλλακτικές μέθοδοι συνδυασμού ταξινομητών. Η Ενότητα 2.4 περιγράφει το πρόβλημα της εξαγωγής πληροφορίας, ενώ γίνεται επισκόπηση και στη διεθνή βιβλιογραφία. Τέλος, η Ενότητα 2.5 περιγράφει τη βιβλιογραφία στην περιοχή του συνδυασμού συστημάτων εξαγωγής πληροφορίας ενώ αιτιολογείται η χρήση τεχνικών ψηφοφορίας και συσσωρευμένης γενίκευσης για το συνδυασμό, όπως προτείνονται σε αυτή τη διατριβή. Το κεφάλαιο αυτό αποτελεί σημαντικό υπόβαθρο για τη μελέτη του υπολοίπου της διατριβής.

2.1 Βασικές έννοιες μηχανικής μάθησης

Η *μηχανική μάθηση (machine learning)* αποσκοπεί γενικά στην κατασκευή ενός υπολογιστικού συστήματος το οποίο θα μπορεί αυτόματα να βελτιώνεται με βάση τα δεδομένα που επεξεργάζεται. Η μηχανική μάθηση αποτελεί σημαντικό κομμάτι του κλάδου της *τεχνητής νοημοσύνης (artificial intelligence)*, καθώς η δυνατότητα μάθησης αποτελεί βασικό χαρακτηριστικό κάθε ευφυούς (νοήμονος) υπολογιστικού συστήματος.

Η χρήση της μηχανικής μάθησης για την απόκτηση νέας γνώσης μπορεί να θεωρηθεί ως ένα πρόβλημα αναζήτησης σε ένα χώρο πιθανών υποθέσεων, εκείνης που ταιριάζει καλύτερα στα δεδομένα εκπαίδευσης. Ως ένα πρόβλημα αναζήτησης πλέον, οι παράμετροι που υπεισέρχονται σε ένα πρόβλημα μηχανικής μάθησης είναι παρόμοιες με εκείνες των αλγορίθμων αναζήτησης. Για παράδειγμα, απαιτείται μια κατάλληλη

γλώσσα αναπαράστασης του χώρου των πιθανών υποθέσεων, όπως ο *κατηγορηματικός λογισμός (predicate calculus)*, η χρήση *τελεστών (operators)* για τις μεταβάσεις ανάμεσα στις υποθέσεις, καθώς και μια μετρική αξιολόγησης κάθε υπόθεσης.

Η αναζήτηση της υπόθεσης εκείνης που ταιριάζει καλύτερα στα παραδείγματα εκπαίδευσης αποτελεί τυπικό παράδειγμα *επαγωγικής μάθησης (inductive learning)*. Δοθέντων των παραδειγμάτων “Το φαγητό της Κορέας είναι πικάντικο”, “Το φαγητό της Ιαπωνίας είναι πικάντικο”, ”Το φαγητό της Γερμανίας δεν είναι πικάντικο”, τότε η έξοδος του συστήματος μάθησης θα μπορούσε να είναι η υπόθεση “Το φαγητό της Ασίας είναι πικάντικο”. Για να παραχθεί ένα τέτοιο συμπέρασμα, θα πρέπει το σύστημα μάθησης να διαθέτει την πληροφορία ότι η Κορέα και η Ιαπωνία είναι χώρες της Ασίας.

Το παραπάνω παράδειγμα αποτελεί παράδειγμα *μάθησης εννοιών (concept learning)*, όπου η *έννοια στόχος (target concept)* είναι η έννοια “χώρα με πικάντικο φαγητό”. Κάθε *έννοια στόχος* μπορεί να θεωρηθεί ως μια λογική συνάρτηση που αποφαίνεται θετικά για όσα παραδείγματα εκπαίδευσης αποτελούν παραδείγματα της έννοιας και αρνητικά για την αντίθετη περίπτωση. Τα φαγητά της Κορέας και της Ιαπωνίας αποτελούν θετικά παραδείγματα για την έννοια “χώρα με πικάντικο φαγητό”, ενώ το φαγητό της Γερμανίας αποτελεί αρνητικό παράδειγμα. Η υπόθεση που αναζητείται *με βάση ένα συγκεκριμένο σύνολο δεδομένων εκπαίδευσης* θα πρέπει να συμβαδίζει με την *έννοια στόχο* κάτι που ισχύει στο παραπάνω παράδειγμα.

Το κυρίαρχο πρόβλημα της επαγωγικής μάθησης είναι ότι πολλές υποθέσεις μπορεί να ταιριάζουν με τα δεδομένα εκπαίδευσης. Στο προηγούμενο παράδειγμα, η υπόθεση “το φαγητό της Κορέας ή της Ιαπωνίας είναι πικάντικο” ταιριάζει επίσης στα δεδομένα. Εδώ εισάγεται η έννοια της *επαγωγικής κλίσης (inductive bias)*, η οποία αναφέρεται σε ένα σύνολο περιορισμών για την υπόθεση που ταιριάζει καλύτερα στα δεδομένα εκπαίδευσης. Στο παράδειγμά μας, η υπόθεση “το φαγητό της Κορέας *ή* της Ιαπωνίας είναι πικάντικο” δεν θα ταιριάζει πλέον στα δεδομένα εκπαίδευσης, εάν τεθεί ένας περιορισμός που να υποδεικνύει ότι η υπόθεση που αναζητείται δε μπορεί να αποτελεί διάζευξη άλλων υποθέσεων και συγκεκριμένα των υποθέσεων “ το φαγητό της Κορέας είναι πικάντικο” και “το φαγητό της Ιαπωνίας είναι πικάντικο”.

Ένα κυρίαρχο χαρακτηριστικό της μάθησης είναι ότι η υπόθεση που ταιριάζει καλύτερα στα παραδείγματα εκπαίδευσης θα πρέπει να είναι αρκετά γενική από τα παραδείγματα εκπαίδευσης ώστε να προσεγγίζει την *έννοια στόχο* και σε παραδείγματα άγνωστα κατά τη μάθηση. Η υπόθεση “Τα φαγητό της Ασίας είναι πικάντικο” μπορεί να αποφανθεί

θετικά για το παράδειγμα “Το φαγητό της Ταϊλάνδης είναι πικάντικο” που δεν υπήρχε στο αρχικό σύνολο παραδειγμάτων εκπαίδευσης. Αντίθετα η υπόθεση “το φαγητό της Κορέας ή της Ιαπωνίας είναι πικάντικο” δεν είναι αρκετά γενική για να αποφανθεί σωστά για το τελευταίο παράδειγμα.

Η μηχανική μάθηση μπορεί να διακριθεί στην *επιβλεπόμενη μάθηση (supervised learning)* και στη *μάθηση χωρίς επίβλεψη (unsupervised learning)*. Τυπικό παράδειγμα επιβλεπόμενης μάθησης είναι τα προβλήματα *ταξινόμησης (classification)*. Σε ένα πρόβλημα ταξινόμησης, κάθε παράδειγμα εκπαίδευσης αντιστοιχεί σε ένα διάνυσμα $< x_1 ... x_n , y >$, όπου $x_1 ... x_n$ είναι ένα σύνολο τιμών χαρακτηριστικών, ή αλλιώς γνωρισμάτων, και y είναι μια τιμή κλάσης η οποία περιγράφει ένα συγκεκριμένο γεγονός για μια θεματική περιοχή, ή αλλιώς, την *έννοια στόχο*. Για το απλοποιημένο παράδειγμα με τα πικάντικα φαγητά, κάθε παράδειγμα εκπαίδευσης αντιστοιχεί σε ένα διάνυσμα χαρακτηριστικών $< x, y >$, όπου το x είναι το όνομα μιας χώρας, ενώ το y παίρνει τις τιμές *ναι* και *όχι*, ανάλογα με το αν τα φαγητά της συγκεκριμένης χώρας είναι πικάντικα ή όχι. Η τιμή y πρέπει να προβλεφθεί από το σύστημα μάθησης κατά τη διαδικασία αξιολόγησης σε ένα άγνωστο παράδειγμα $< x_1 ... x_n >$.

Στη μάθηση χωρίς επίβλεψη, δεν υπάρχει προκαθορισμένο σύνολο τιμών. Τα παραδείγματα εκπαίδευσης χωρίζονται σε, άγνωστες εκ των προτέρων, ομάδες με βάση τα χαρακτηριστικά τους. Παραδείγματα αλγορίθμων μη επιβλεπόμενης μάθησης αποτελούν οι αλγόριθμοι *COBWEB* [44], *Apriori* [3], *AutoClass* [22] κ.α. Χαρακτηριστικό παράδειγμα μη επιβλεπόμενης μάθησης αποτελεί η εύρεση *κανόνων συσχέτισης (association rules)* της μορφής “εάν Χ τότε Υ”, όπου Χ και Υ είναι τιμές που συνδέουν τιμές χαρακτηριστικών στα διανύσματα εκπαίδευσης. Ένας τέτοιος κανόνας θα μπορούσε να λέει “οι πελάτες που αγοράζουν μπύρα, αγοράζουν και ξηρούς καρπούς” κατά την ανάλυση των προϊόντων που αγοράζονται από πελάτες σε ένα κατάστημα. Αυτή η διατριβή ασχολείται με προβλήματα επιβλεπόμενης μηχανικής μάθησης.

Πληθώρα αλγορίθμων μηχανικής μάθησης, από την άλλη πλευρά, είναι σχεδιασμένοι για προβλήματα ταξινόμησης, όπως είναι οι αλγόριθμοι *ID3* [89] και *C4.5* [90] για την εκμάθηση *δέντρων απόφασης (decision trees)*, ο αλγόριθμος *Naïve Bayes* [60], ο αλγόριθμος των *κ-κοντινότερων γειτόνων* (*k-nearest-neighbors*, [4]) κ.α. Το εκπαιδευμένο μοντέλο που προκύπτει από την εφαρμογή ενός αλγορίθμου ταξινόμησης σε ένα σύνολο διανυσμάτων χαρακτηριστικών καλείται και *ταξινομητής (classifier)*.

Παρόλο που η διαδικασία μάθησης στους υπολογιστές απέχει αρκετά από τη διαδικασία μάθησης στους ανθρώπους, πληθώρα εφαρμογών έχουν επιτυχώς αναπτυχθεί τα

τελευταία χρόνια οι οποίες χρησιμοποιούν τη μηχανική μάθηση σε διάφορους τομείς όπως ο αυτόματος εντοπισμός κάλπικων συναλλαγών με πιστωτικές κάρτες, η *ανακάλυψη γνώσης (knowledge discovery)* από μεγάλο όγκο βάσεων δεδομένων, η αυτόματη οδήγηση οχημάτων σε μεγάλες λεωφόρους, καθώς και η εκμάθηση επιτραπέζιων παιχνιδιών χειριζόμενων από τον υπολογιστή, όπως τάβλι και σκάκι, και μάλιστα σε επίπεδο συγκρίσιμο με εκείνο των παγκόσμιων πρωταθλητών.

Στο χώρο του παγκοσμίου ιστού, η μηχανική μάθηση έχει αποκτήσει μεγάλο εύρος εφαρμογής τα τελευταία χρόνια. Κύριος στόχος είναι η αντιμετώπιση του προβλήματος της *υπερ-πληροφόρησης (information overload)* μέσω της ανάπτυξης εφαρμογών οι οποίες θα μπορούν αυτόματα να φιλτράρουν τον τεράστιο όγκο δεδομένων του ιστού, αναζητώντας σχετική πληροφορία για τον τελικό χρήστη. Τεχνικές *ανάκτησης πληροφορίας (information retrieval)* βασίζονται αρκετά στη χρήση μηχανικής μάθησης για την ανάκτηση και κατηγοριοποίηση σελίδων του ιστού που είναι σχετικές για τον χρήστη. Πιο πρόσφατα, τεχνικές *εξαγωγής πληροφορίας (information extraction)* χρησιμοποιούν επίσης μηχανική μάθηση για τον εντοπισμό συγκεκριμένης πληροφορίας μέσα στα κείμενα που είναι σχετικά με τα ενδιαφέροντα του χρήστη.

Στόχος της ενότητας αυτής ήταν να παράσχει κάποιες πολύ βασικές έννοιες που αφορούν τη μηχανική μάθηση. Μια περισσότερο λεπτομερής περιγραφή του πεδίου της μηχανικής μάθησης υπάρχει στο βιβλίο [80].

2.2 Συνδυασμός ταξινομητών με χρήση ψηφοφορίας

Το μεγαλύτερο τμήμα της ερευνητικής δραστηριότητας στο χώρο της μηχανικής μάθησης αφορά την *επιβλεπόμενη μάθηση*, τυπικό παράδειγμα της οποίας είναι τα προβλήματα *ταξινόμησης*. Το κίνητρο για το συνδυασμό ταξινομητών οφείλεται κυρίως στην παρατήρηση ότι δεν είναι δυνατό να βρεθεί ένας ταξινομητής που να είναι ο καλύτερος σε όλες τις θεματικές περιοχές, κάτι που είναι γνωστό και ως “no free lunch theorem” [121] ή “conservation law of generalization performance” [94]. Από την άλλη μεριά, ο συνδυασμός ταξινομητών μπορεί να οδηγήσει στην επίτευξη καλύτερων αποτελεσμάτων σε μετα-επίπεδο. Ο απλούστερος τρόπος συνδυασμού ταξινομητών σε μετα-επίπεδο είναι με χρήση τεχνικών *ψηφοφορίας*.

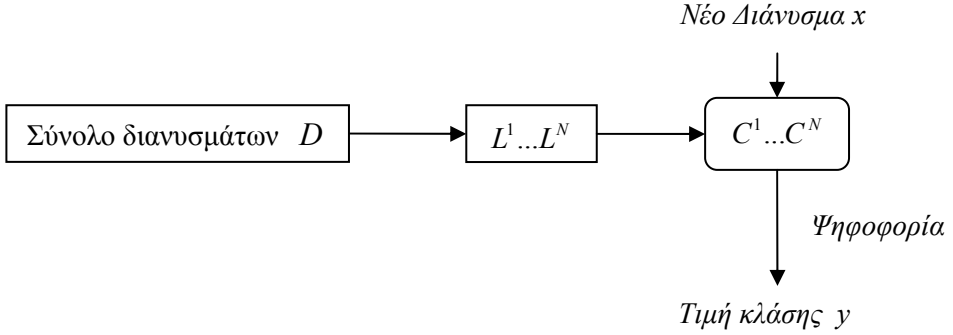
2.2.1 Ορισμός

Έστω $C^1 ... C^N$ ένα σύνολο ταξινομητών που προκύπτουν από την εφαρμογή N διαφορετικών αλγορίθμων μηχανικής μάθησης $L^1 ... L^N$ σε ένα σύνολο D από

διανύσματα χαρακτηριστικών. Για την ταξινόμηση ενός νέου παραδείγματος (διανύσματος) κατά τη *διαδικασία επαλήθευσης (runtime)*, οι εκπαιδευμένοι ταξινομητές $C^1 ... C^N$ σε μετα-επίπεδο προβλέπουν μια ονομαστική τιμή ως χαρακτηριστικό κλάσης του νέου διανύσματος. Η τιμή με τον μεγαλύτερο αριθμό ψήφων επιλέγεται τελικά ως χαρακτηριστικό κλάσης του νέου διανύσματος. Η μέθοδος αυτή είναι γνωστή και ως *πλειοψηφική ψηφοφορία (plurality or majority voting)* και είναι ο απλούστερος τρόπος συνδυασμού της εξόδου πολλαπλών ταξινομητών. Τυπικά, οι όροι *plurality voting* και *majority voting* διαφέρουν, καθώς στην τελευταία περίπτωση οι μισοί τουλάχιστον ψήφοι θα πρέπει να ανήκουν στη νικήτρια τιμή για το χαρακτηριστικό κλάσης.

Διάφορες επεκτάσεις της πλειοψηφικής ψηφοφορίας για το συνδυασμό πολλαπλών ταξινομητών υπάρχουν στη βιβλιογραφία, οι οποίες περιλαμβάνουν τη *σταθμισμένη πλειοψηφική ψηφοφορία (weighted majority voting)*, και την *ψηφοφορία με χρήση πιθανοτικών κατανομών (voting with probability distributions)*. Στην πρώτη περίπτωση, η ψήφος κάθε ταξινομητή σταθμίζεται ανάλογα με την ακρίβειά του, όπως αυτή αξιολογείται σε ένα ξεχωριστό τμήμα δεδομένων εκπαίδευσης, ή με διαδικασία *διασταυρωμένης επικύρωσης (cross-validation)* στα δεδομένα. Στην δεύτερη περίπτωση, η έξοδος κάθε ταξινομητή είναι μια πιθανοτική κατανομή σε όλες τις κλάσεις. Για κάθε κλάση, αθροίζονται οι αντίστοιχες πιθανότητες (ή υπολογίζεται ο μέσος όρος τους) από τους N ταξινομητές και η κλάση με το μεγαλύτερο άθροισμα (ή μέσο όρο) επιλέγεται τελικά.

Το Σχήμα 2.1 δείχνει σχηματικά πως λειτουργεί η ψηφοφορία πολλαπλών ταξινομητών, δοθέντος ενός διανύσματος κατά τη διαδικασία επαλήθευσης. Η έξοδος κάθε ταξινομητή για το νέο διάνυσμα μπορεί να είναι είτε μια ονομαστική τιμή είτε μια πιθανοτική κατανομή σε όλες τις σχετικές κλάσεις για μια θεματική περιοχή.



Σχήμα 2.1. Συνδυασμός διαφορετικών ταξινομητών με χρήση ψηφοφορίας.

2.2.2 Διεθνής επισκόπηση

Η θεωρία *Dempster-Shafer*, είναι επίσης μια αρκετά γνωστή και πανίσχυρη θεωρία, βάσει της οποίας υπολογίζεται μια συνδυασμένη πιθανότητα για ένα συγκεκριμένο γεγονός από τις προβλέψεις πολλαπλών ταξινομητών, λαμβάνοντας υπόψη την *αβεβαιότητα (uncertainty)* στις προβλέψεις των ταξινομητών. Τέλος, ψηφοφορία λαμβάνει χώρα με χρήση των υπολογισμένων πιθανοτήτων. Λεπτομέρειες για τη θεωρία *Dempster-Shafer* μπορούν να βρεθούν στις εργασίες [6, 101].

Πρέπει να τονιστεί ότι μέθοδοι όπως η *ενδυνάμωση (boosting,* [51]*)* και η *εμφωλίαση (bagging,* [13]*)* χρησιμοποιούν επίσης ψηφοφορία σε ένα σύνολο $C^1 ... C^N$ ταξινομητών που παράγονται όμως από την εφαρμογή του ιδίου αλγορίθμου σε N διαφορετικές εκδόσεις του συνόλου των διανυσμάτων εκπαίδευσης. Στη μέθοδο της εμφωλίασης, οι N διαφορετικές εκδόσεις του συνόλου εκπαίδευσης, δηλαδή των *διανυσμάτων χαρακτηριστικών (feature vectors)*, δημιουργούνται με *τυχαία δειγματοληψία και επανατοποθέτηση (random sampling with replacement)* ή αλλιώς, *αυτοδύναμη δειγματοληψία (bootstrap sampling)*. Στη μέθοδο της ενδυνάμωσης δεν πραγματοποιείται δειγματοληψία στα διανύσματα εκπαίδευσης, όπως στη μέθοδο της εμφωλίασης, αλλά σε κάθε γύρο γίνεται επαναπροσδιορισμός των βαρών των διανυσμάτων που έχουν λανθασμένα ταξινομηθεί.

Η δημιουργία των διαφορετικών εκδόσεων των δεδομένων εκπαίδευσης μπορεί να προκύψει όχι μόνο με δειγματοληψία κι επανατοποθέτηση (όπως στην εμφωλίαση) ή με επαναπροσδιορισμό των βαρών (όπως στην ενδυνάμωση), αλλά και με δειγματοληψία στα χαρακτηριστικά των διανυσμάτων εκπαίδευσης. Στην εργασία [23] δημιουργούνται χειρονακτικά 8 ανισομεγέθη υποσύνολα των 119 συνολικά διαθέσιμων χαρακτηριστικών, οδηγώντας σε 8 διαφορετικά σύνολα εκπαίδευσης στα οποία εφαρμόζεται ένας αλγόριθμος επαγωγής νευρωνικών δικτύων. Σύμφωνα με τον Dietterich [37], όμως, η προσέγγιση αυτή μπορεί να είναι αποτελεσματική μόνο όταν ένα μεγάλο μέρος των χαρακτηριστικών στα διανύσματα είναι *πλεονάζοντα (redundant)*.

Οι ταξινομητές που παράγονται με την εφαρμογή ενός και μόνο αλγορίθμου σε διαφορετικές εκδόσεις των δεδομένων εκπαίδευσης (όπως για παράδειγμα στις μεθόδους της ενδυνάμωσης και εμφωλίασης) και ως *ομογενείς ταξινομητές (homogenous classifiers)*. Αντίθετα, οι ταξινομητές που προκύπτουν από την εφαρμογή N διαφορετικών αλγορίθμων στο ίδιο σύνολο διανυσμάτων, αναφέρονται και ως *ετερογενείς ταξινομητές (heterogeneous classifiers)*. Στην εργασία [112] παρουσιάζεται μια απλή, αλλά ταυτόχρονα αποδοτική, μέθοδος ψηφοφορίας ετερογενών ταξινομητών.

Σύμφωνα με τη μέθοδο αυτή, πραγματοποιείται ψηφοφορία μόνο μεταξύ των ταξινομητών εκείνων των οποίων οι διαφορές στις προβλέψεις είναι στατιστικά σημαντικές, σύμφωνα με το γνωστό τεστ *paired t-test* [38].

Τέλος, υπάρχει πλούσια βιβλιογραφία όσον αφορά το συνδυασμό ομογενών ταξινομητών, μέσω ψηφοφορίας, που προκύπτουν από την εφαρμογή αλγορίθμων όπως τα νευρωνικά δίκτυα [23, 67, 86], τα δέντρα απόφασης [72, 85], οι κανόνες επαγωγής [65] κ.α. Οι εκπαιδευμένοι ταξινομητές (ομογενείς ή ετερογενείς), μαζί με τη μεθοδολογία συνδυασμού τους σε μετα-επίπεδο, καλούνται και *συγκρότημα ταξινομητών* (*ensemble of classifiers*, [37]). Εκτενείς παρουσιάσεις για θέματα ψηφοφορίας ταξινομητών μπορούν να βρεθούν επίσης στις εργασίες [9, 37].

2.3 Συνδυασμός ταξινομητών με χρήση συσσωρευμένης γενίκευσης

Αντί της πραγματοποίησης απλής ψηφοφορίας, μεγαλύτερο ενδιαφέρον παρουσιάζει η χρήση μηχανικής μάθησης σε μετα-επίπεδο για το συνδυασμό ταξινομητών.

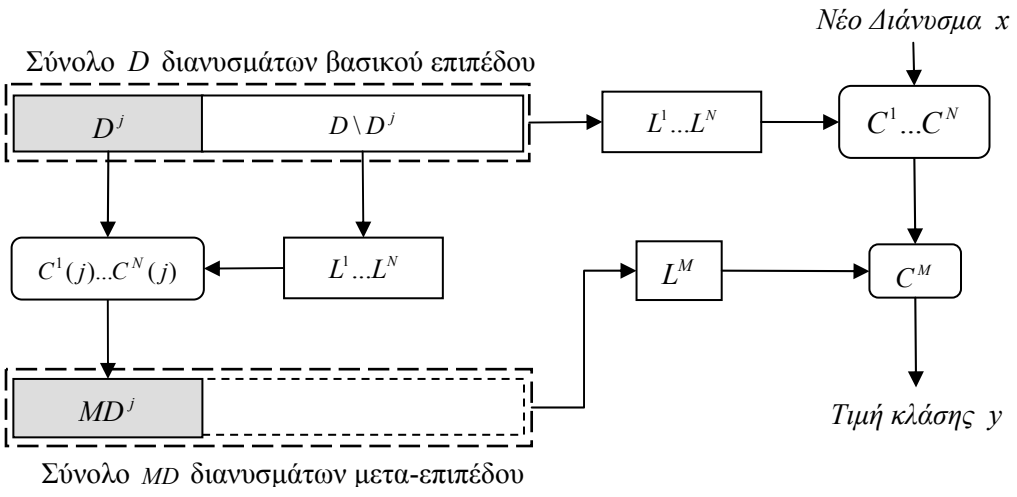
2.3.1 Ορισμός

Ο Wolpert [120] όρισε μια καινοτομική προσέγγιση για το συνδυασμό ταξινομητών με χρήση μηχανικής μάθησης, την οποία ονόμασε *συσσωρευμένη γενίκευση (stacked generalization)* ή *συσσώρευση (stacking)*. Η βασική ιδέα πίσω από τη συσσωρευμένη γενίκευση είναι η εκπαίδευση ενός ταξινομητή σε μετα-επίπεδο (επίπεδο-1) πάνω σε διανύσματα χαρακτηριστικών που κατασκευάζονται από την έξοδο ενός συνόλου ταξινομητών σε βασικό επίπεδο (επίπεδο-0) και μέσω μιας διαδικασίας *διασταυρωμένης επικύρωσης (cross-validation)*, η οποία έχει ως εξής:

Έστω D ένα σύνολο δεδομένων εκπαίδευσης το οποίο αποτελείται από διανύσματα χαρακτηριστικών. Το σύνολο D αναφέρεται και ως σύνολο δεδομένων βασικού επιπέδου (ή επιπέδου-0). Έστω επίσης $L^1 ... L^N$ ένα σύνολο N διαφορετικών αλγορίθμων μηχανικής μάθησης, σχεδιασμένων για κοινά προβλήματα ταξινόμησης. Οι αλγόριθμοι $L^1 ... L^N$ καλούνται επίσης αλγόριθμοι βασικού επιπέδου (ή επιπέδου-0). Κατά τη διάρκεια μιας διαδικασίας διασταυρωμένης επικύρωσης J βημάτων, το σύνολο D των διανυσμάτων εκπαίδευσης χωρίζεται με τυχαία δειγματοληψία σε J διακεκριμένα τμήματα $D^1 ... D^J$ σχεδόν ίδιου μήκους. Σε κάθε *j*-βήμα της διαδικασίας, $j = 1..J$, οι αλγόριθμοι $L^1 ... L^N$ μηχανικής μάθησης εφαρμόζονται στο υποσύνολο $D \setminus D^j$ των διανυσμάτων εκπαίδευσης και οι εκπαιδευμένοι ταξινομητές $C^1(j) ... C^N(j)$ εφαρμόζονται στο υποσύνολο D^j των διανυσμάτων. Για κάθε διάνυσμα *x* στο σύνολο D^j, οι προβλέψεις των ταξινομητών ενώνονται μαζί με την πρωτότυπη (επισημειωμένη)

τιμή κλάσης $y(x)$ για το διάνυσμα και σχηματίζουν σε μετα-επίπεδο ένα νέο σύνολο MD^j διανυσμάτων χαρακτηριστικών.

Στο τέλος ολόκληρης της διαδικασίας διασταυρωμένης επικύρωσης, η ένωση $MD = \cup MD^j$, $j = 1..J$ αποτελεί το σύνολο των διανυσμάτων χαρακτηριστικών σε μετα-επίπεδο. Το σύνολο αυτό αναφέρεται αλλιώς και ως σύνολο επιπέδου-1 και θα χρησιμοποιηθεί για την εκπαίδευση ενός ταξινομητή C^M σε μετα-επίπεδο, από την εφαρμογή ενός αλγορίθμου L^M. Ο αλγόριθμος L^M μπορεί να είναι ένας από τους $L^1...L^N$ ή διαφορετικός. Τελικά, οι αλγόριθμοι $L^1...L^N$ εφαρμόζονται ξανά σε ολόκληρο το σύνολο D των διανυσμάτων χαρακτηριστικών, ώστε να προκύψουν οι τελικοί ταξινομητές $C^1...C^N$ οι οποίοι θα χρησιμοποιηθούν κατά τη διαδικασία επαλήθευσης. Για την ταξινόμηση ενός νέου διανύσματος *x* κατά τη διαδικασία επαλήθευσης, οι προβλέψεις από τους εκπαιδευμένους ταξινομητές $C^1...C^N$ του βασικού επιπέδου ενώνονται και σχηματίζουν ένα νέο διάνυσμα για το οποίο ο ταξινομητής C^M του μετα-επιπέδου καλείται να προβλέψει την τελική τιμή για το χαρακτηριστικό κλάσης. Το Σχήμα 2.2(α) δείχνει τη διαδικασία διασταυρωμένης επικύρωσης για τη δημιουργία των διανυσμάτων εκπαίδευσης και των ταξινομητών σε μετα-επίπεδο. Το Σχήμα 2.2(β) επιδεικνύει τη μεθοδολογία συσσωρευμένης γενίκευσης κατά την επαλήθευση.



Σχήμα 2.2 Συνδυασμός πολλαπλών ταξινομητών με χρήση συσσωρευμένης γενίκευσης.

Ο Wolpert όρισε τη συσσωρευμένη γενίκευση ακόμα και για την απλή περίπτωση του ενός ταξινομητή σε βασικό επίπεδο. Στη βιβλιογραφία όμως έχει επικρατήσει η περίπτωση της ύπαρξης πολλαπλών ταξινομητών σε βασικό επίπεδο. Η ιδέα για τον ορισμό της συσσωρεμένης γενίκευσης ξεκίνησε από τον προβληματισμό για το ποιος ταξινομητής είναι καλύτερος για ένα σύνολο δεδομένων. Η επικρατέστερη προσέγγιση

για την επιλογή ταξινομητών εμπειρικά, είναι η πραγματοποίηση διασταυρωμένης επικύρωσης στα διανύσματα εκπαίδευσης για την εκπαίδευση και αξιολόγηση πολλαπλών ταξινομητών και κατόπιν η επιλογή εκείνου με το μικρότερο *λάθος γενίκευσης (generalization error)*.

Η πρόταση του Wolpert ήταν ότι από τη στιγμή που πραγματοποιείται διασταυρωμένη επικύρωση στα δεδομένα εκπαίδευσης, η έξοδος των πολλαπλών ταξινομητών θα μπορούσε να χρησιμοποιηθεί για την κατασκευή ενός δεύτερου συνόλου διανυσμάτων (με κοινά χαρακτηριστικά κλάσης με τα διανύσματα του αρχικού συνόλου). Το νέο αυτό σύνολο διανυσμάτων θα μπορούσε στη συνέχεια να χρησιμοποιηθεί για την εκπαίδευση ενός άλλου ταξινομητή σε μετα-επίπεδο με στόχο μικρότερο λάθος γενίκευσης από όλους τους ταξινομητές του βασικού επιπέδου.

Ο Breiman [14], μελέτησε τη συσσώρευση πολλαπλών αλγορίθμων μηχανικής μάθησης, αλλά για προβλήματα *αριθμητικής πρόβλεψης (numeric prediction)* και όχι για προβλήματα ταξινόμησης, όπως ο Wolpert [120]. Στην αριθμητική πρόβλεψη, το χαρακτηριστικό κλάσης στα διανύσματα δεν παίρνει τιμές από ένα διακριτό σύνολο, αλλά είναι μια αριθμητική τιμή. Ο Breiman χρησιμοποίησε τον όρο *συσσωρευμένη παλινδρόμηση (stacked regression)*, εξαιτίας και του γεγονότος ότι όλοι οι αλγόριθμοι σε βασικό και σε μετα-επίπεδο είναι αλγόριθμοι παλινδρόμησης.

Η συσσωρευμένη γενίκευση τυπικά επιτυγχάνει καλύτερα αποτελέσματα από την ψηφοφορία, όσον αφορά το συνδυασμό πολλαπλών αλγορίθμων. Η ψηφοφορία, όμως, είναι άμεσα εφαρμόσιμη κατά τη διαδικασία επαλήθευσης, ενώ η συσσωρευμένη γενίκευση απαιτεί διασταυρωμένη επικύρωση στα διανύσματα εκπαίδευσης, όπως φαίνεται και από τη σύγκριση των Σχημάτων 2.1 και 2.2.

2.3.2 Διεθνής επισκόπηση

Η έρευνα στη συσσωρευμένη γενίκευση αφορά δύο σημαντικά ζητήματα, τα οποία χαρακτηρίστηκαν ως *μαύρη τέχνη (black art)* από τον Wolpert [120]. Το πρώτο ζήτημα αφορά την επιλογή των ταξινομητών, τόσο σε βασικό όσο και σε μετα-επίπεδο, η οποία θα οδηγήσει στα καλύτερα εμπειρικά αποτελέσματα. Το δεύτερο ζήτημα, το οποίο έχει γενικά λάβει μεγαλύτερη προσοχή από τους ερευνητές, αφορά το συνδυασμό των προβλέψεων των ταξινομητών του βασικού επιπέδου και την αντιστοιχία τους σε διανύσματα χαρακτηριστικών στο μετα-επίπεδο. Διατυπωμένο διαφορετικά, το ζητούμενο είναι ποια διανυσματική αναπαράσταση σε μετα-επίπεδο μπορεί να οδηγήσει

σε καλύτερα αποτελέσματα. Τυπικά χαρακτηριστικά που χρησιμοποιούνται είναι οι ονομαστικές τιμές κλάσης των ταξινομητών του βασικού επιπέδου.

Οι Chan και Stolfo [19] πειραματίστηκαν με διάφορες αναπαραστάσεις στα διανύσματα του μετα-επιπέδου, συμπεριλαμβανομένης και της αναπαράστασης με διακριτικό τίτλο *μετα-χαρακτηριστικό-κλάσης (meta-class-attribute)*, όπου στις ονομαστικές τιμές κλάσης που προβλέπονται από τους ταξινομητές του βασικού επιπέδου, προστίθενται και τα χαρακτηριστικά των διανυσμάτων του βασικού επιπέδου, μαζί ασφαλώς με το σωστό χαρακτηριστικό κλάσης για κάθε διάνυσμα. Πειραματίστηκαν επίσης με ένα σχήμα *αρχιτεκτονικής-διαιτητή (arbiter architecture)*, όπου ένας ταξινομητής σε μετα-επίπεδο εκπαιδεύεται μόνο σε ένα υποσύνολο των διανυσμάτων του βασικού επιπέδου, όπου οι ταξινομητές σε βασικό επίπεδο διαφωνούν στις προβλέψεις τους. Ένα *υβριδικό* σχήμα αξιολογήθηκε επίσης από τους Chan και Stolfo, όπου ένας ταξινομητής σε μετα-επίπεδο εκπαιδεύεται μόνο σε ένα υποσύνολο των διανυσμάτων του μετα-επιπέδου που έχουν κατασκευαστεί με χρήση της αναπαράστασης *μετα-χαρακτηριστικό-κλάσης* και όπου οι ταξινομητές σε βασικό επίπεδο διαφωνούν στις προβλέψεις τους. Τα πειράματα ανέδειξαν την μικρή ανωτερότητα της αναπαράστασης *μετα-χαρακτηριστικό-κλάσης*, αλλά οι διαφορές μετρήθηκαν ως στατιστικά μη σημαντικές. Ο Schaffer [93] αμφισβήτησε τη χρησιμότητα της αναπαράστασης αυτής, η οποία είναι γνωστή και ως *συσσώρευση δύο επιπέδων (bi-level stacking)*.

Η εργασία των Chan και Stolfo [19] είναι επίσης γνωστή και ως *μετα-μάθηση (meta-learning)*. Ο όρος αυτός έχει χρησιμοποιηθεί για να περιγράψει το πρόβλημα της εύρεσης του κατάλληλου αλγορίθμου για ένα συγκεκριμένο σύνολο δεδομένων [12, 87]. Μια άριστη επισκόπηση πάνω σε θέματα μετα-μάθησης υπάρχει στην εργασία [114].

Οι Ting και Witten [107] παρουσίασαν μια παραλλαγή της συσσωρευμένης γενίκευσης όπου κάθε ταξινομητής του βασικού επιπέδου επιστρέφει μια πιθανοτική κατανομή (διάνυσμα) σε όλες τις δυνατές τιμές που παίρνει το χαρακτηριστικό κλάσης, αντί μιας ονομαστικής τιμής κλάσης. Τα ξεχωριστά διανύσματα από τους N ταξινομητές ενώνονται προς ένα ενιαίο διάνυσμα των $N*Q$ χαρακτηριστικών, όπου Q είναι ο αριθμός των σχετικών κλάσεων για μια θεματική περιοχή. Οι Ting και Witten [107] πρότειναν επίσης τη χρήση *πολύ-ανταποκριτικής γραμμικής παλινδρόμησης (multi-response linear regression, MLR)* ως ταξινομητή σε μετα-επίπεδο, η οποία αποδείχτηκε ιδιαίτερα αποτελεσματική σε σύγκριση με άλλους ταξινομητές που χρησιμοποιήθηκαν σε μετα-επίπεδο. Πρέπει να υπενθυμιστεί ότι η χρήση παλινδρόμησης σε μετα-επίπεδο,

δε σημαίνει ότι χειριζόμαστε πλέον ένα πρόβλημα αριθμητικής πρόβλεψης. Απλά το πρόβλημα ταξινόμησης μοντελοποιείται με χρήση αριθμητικής πρόβλεψης.

Συγκεκριμένα, η *πολύ-ανταποκριτική γραμμική παλινδρόμηση* που προτείνουν οι Ting και Witten [107] είναι μια προσαρμογή της γραμμικής παλινδρόμησης [14], κατά την οποία ένα πρόβλημα ταξινόμησης Q κλάσεων μετασχηματίζεται σε Q διαφορετικά προβλήματα *δυαδικής ταξινόμησης (binary classification)*: για κάθε σχετική κλάση, μια γραμμική εξίσωση υπολογίζεται με βάση τα διανύσματα εκπαίδευσης, η οποία επιστρέφει την τιμή ένα, εάν η τιμή του χαρακτηριστικού κλάσης ενός διανύσματος συμπίπτει με την υπό εξέταση κλάση, ή την τιμή μηδέν σε διαφορετική περίπτωση. Δοθέντος ενός νέου διανύσματος για να ταξινομηθεί σε μετα-επίπεδο, για κάθε σχετική κλάση επιστρέφεται μια τιμή από την αντίστοιχη εξίσωση γραμμικής παλινδρόμησης και η κλάση με τη μεγαλύτερη τιμή επιλέγεται για το νέο διάνυσμα.

Ο Merz [78] πρότεινε μια μέθοδο συσσωρευμένης γενίκευσης, την οποία ονόμασε SCANN, κατά την οποία εφαρμόζεται σε μετα-επίπεδο η θεωρία *ανάλυσης αντιστοιχίας (correspondence analysis)* για τον εντοπισμό των συσχετίσεων μεταξύ των προβλέψεων των ταξινομητών του βασικού επιπέδου. Οι συσχετίσεις που εντοπίζονται απομακρύνονται από το σύνολο των διανυσμάτων χαρακτηριστικών του μετα-επιπέδου κι ένας αλγόριθμος κοντινότερου γείτονα εφαρμόζεται στη συνέχεια.

Οι Seewald & Fürnkranz [99] αξιολόγησαν μια μέθοδο *βαθμολόγησης ταξινομητών (grading classifiers)*, όπου ένας ταξινομητής εκπαιδεύεται σε μετα-επίπεδο για κάθε αντίστοιχο του βασικού επιπέδου. Ο ταξινομητής του μετα-επιπέδου μαθαίνει πότε ο αντίστοιχος του βασικού κάνει σωστές προβλέψεις, χρησιμοποιώντας το σύνολο των διανυσμάτων χαρακτηριστικών του βασικού επιπέδου. Στο τέλος ψηφοφορία με χρήση βαρών πραγματοποιείται ανάμεσα στους ταξινομητές του μετα-επιπέδου για την επιλογή της τελικής τιμής κλάσης. Η αξιολόγηση όμως δεν οδήγησε σε καλύτερα αποτελέσματα από τη συσσωρευμένη γενίκευση.

Οι Todorovski και Džeroski [110] πρότειναν μια μέθοδο συσσωρευμένης γενίκευσης η οποία βασίζεται στη χρήση, σε μετα-επίπεδο, *δέντρων μετα-απόφασης (meta decision trees, MDTs)*. Τα δέντρα μετα-απόφασης έχουν την ίδια δομή με τα κανονικά δέντρα απόφασης, απλά τα φύλλα τους περιέχουν τους ταξινομητές του βασικού επιπέδου, αντί για προβλέψεις κλάσεων. Επομένως τα δέντρα μετα-απόφασης προσδιορίζουν ποιους ταξινομητές του βασικού επιπέδου πρέπει να εμπιστευτούμε στις προβλέψεις τους. Ως χαρακτηριστικά σε μετα-επίπεδο χρησιμοποιούνται ιδιότητες (όπως η εντροπία και η

μέγιστη πιθανότητα) των πιθανοτήτων κατανομής σε όλες τις κλάσεις, όπως επιστρέφονται από τους ταξινομητές του βασικού επιπέδου.

Οι Todorovski και Džeroski [110] ανέφεραν ότι συσσωρευμένη γενίκευση με χρήση δέντρων μετα-απόφασης επιτυγχάνει καλύτερα αποτελέσματα από την ψηφοφορία, τη συσσωρευμένη γενίκευση με χρήση κανονικών δέντρων απόφασης, καθώς και από τη χρήση *ενδυνάμωσης* ή *εμφωλίασης* κανονικών δέντρων απόφασης. Από την άλλη πλευρά, η χρήση δέντρων μετα-απόφασης επιτυγχάνει μόνο ελαφρώς καλύτερα αποτελέσματα σε σχέση με το σύστημα SCANN που κάνει χρήση *ανάλυσης αντιστοιχίας* και ελαφρώς χειρότερα αποτελέσματα από τη συσσωρευμένη γενίκευση με χρήση *πολύ-ανταποκριτικής γραμμικής παλινδρόμησης*.

Ο Seewald [98] πρότεινε μια τροποποίηση της προσέγγισης από τους Ting και Witten [107], όπου διαφορετικά χαρακτηριστικά χρησιμοποιούνται στα διανύσματα σε μετα-επίπεδο για καθένα από τα Q προβλήματα δυαδικής πρόβλεψης στα οποία έχει μετασχηματιστεί το πρόβλημα ταξινόμησης Q κλάσεων. Πιο συγκεκριμένα, μόνο οι πιθανότητες για κάθε κλάση υπό εξέταση θα πρέπει να χρησιμοποιούνται από κάθε ταξινομητή του βασικού επιπέδου ως χαρακτηριστικά στα διανύσματα του μετα-επιπέδου, αντί να ενώνονται οι πιθανοτικές κατανομές για όλες τις σχετικές κλάσεις από όλους τους ταξινομητές. Ως αποτέλεσμα, ο συνολικός αριθμός των χαρακτηριστικών στα διανύσματα του μετα-επιπέδου μειώνεται από $N*Q$ σε N, μειώνοντας με αυτό τον τρόπο τον χώρο διάστασης του προβλήματος και αυξάνοντας παράλληλα την ταχύτητα της συσσωρευμένης γενίκευσης. Τα πειραματικά αποτελέσματα έδειξαν βελτιωμένα αποτελέσματα, πάλι χρησιμοποιώντας *πολύ-ανταποκριτική γραμμική παλινδρόμηση*, σε σχέση με τα διανύσματα των $N*Q$ χαρακτηριστικών. Η βελτίωση που επιτυγχάνεται αφορά περισσότερο προβλήματα ταξινόμησης *πολλαπλών κλάσεων (multi-class)*.

Οι Džeroski και Ženko (2004) προβληματίστηκαν για το γεγονός ότι η χρήση ψηφοφορίας για το συνδυασμό ταξινομητών μπορεί να αποδειχτεί εξίσου καλή με τη χρήση συσσωρευμένης γενίκευσης σε αρκετά προβλήματα ταξινόμησης και επομένως προτιμητέα αφού είναι υπολογιστικά πιο συμφέρουσα. Για την ανάδειξη της υπεροχής της συσσωρευμένης γενίκευσης, έναντι της ψηφοφορίας, οι Džeroski και Ženko [41] πρότειναν τη χρήση *πολύ-ανταποκριτικών μοντέλων δέντρων (multi-response model trees)* σε μετα-επίπεδο και χρησιμοποιώντας τα διανύσματα πιθανοτικών κατανομών σε όλες τις κλάσεις, όπως αυτά επιστρέφονται από τους ταξινομητές του βασικού επιπέδου. Τα αποτελέσματα όντως επιβεβαίωσαν την υπεροχή της προσέγγισης που

χρησιμοποιεί τα πολύ-ανταποκριτικά μοντέλα δέντρων, σε σύγκριση με τη χρήση πολύ-ανταποκριτικής γραμμικής παλινδρόμησης.

Η χρήση της τελευταίας, μελετήθηκε επίσης από τους Džeroski και Ženko [41] αλλά σε ένα διευρυμένο σύνολο χαρακτηριστικών στα διανύσματα του μετα-επιπέδου, το οποίο αποτελείται από τις πιθανοτικές κατανομές των ταξινομητών του βασικού επιπέδου, επαυξημένο με τις ίδιες κατανομές, πολλαπλασιασμένες με τη μέγιστη πιθανότητα κλάσης, καθώς και με την εντροπία κάθε κατανομής. Ο συνολικός αριθμός χαρακτηριστικών των διανυσμάτων του μετα-επιπέδου θα είναι στην περίπτωση αυτή $N*(2*Q+1)$. Η προσέγγιση αυτή οδήγησε σε καλύτερα αποτελέσματα, σε σύγκριση με τη χρήση μόνο των πιθανοτικών κατανομών ως χαρακτηριστικά διανυσμάτων. Όμως δεν οδήγησε σε καλύτερα αποτελέσματα σε σχέση με τη χρήση πολύ-ανταποκριτικών μοντέλων δέντρων σε μετα-επίπεδο και χρησιμοποιώντας τις πιθανοτικές κατανομές, δηλαδή τα διανύσματα των $N*Q$ χαρακτηριστικών.

Πρόσφατα, τέλος, [18] άρχισε να μελετάται η εφαρμογή της συσσωρευμένης γενίκευσης για προβλήματα *ακολουθιακής μάθησης (sequential learning)*, η οποία καλείται και *συσσωρευμένη ακολουθιακή μάθηση (stacked sequential learning)*. Το κίνητρο για τη χρήση ακολουθιακής μάθησης πηγάζει από το γεγονός ότι κατά την ταξινόμηση ενός διανύσματος, μπορούμε να εκμεταλλευτούμε πληροφορία που βρίσκεται γειτονικά διανύσματα. Κατά τη συσσωρευμένη ακολουθιακή μάθηση, η διασταυρωμένη επικύρωση γίνεται με χρήση ενός αλγορίθμου ταξινόμησης. Κάθε διάνυσμα σε μετα-επίπεδο διατηρεί μεν τα χαρακτηριστικά του αντίστοιχου σε βασικό επίπεδο, αλλά επαυξάνεται με ένα σύνολο $h+f$ χαρακτηριστικών που αντιστοιχούν στις προβλέψεις του ταξινομητή στα προηγούμενα h κι επόμενα f διανύσματα, συν βεβαίως την πρόβλεψη του ταξινομητή για το τρέχον διάνυσμα. Στα νέα αυτά διανύσματα εκπαιδεύεται ένας ταξινομητής σε μετα-επίπεδο. Κατά την επαλήθευση δίνεται μια ακολουθία διανυσμάτων για ταξινόμηση. Μια αντίστοιχη ακολουθία διανυσμάτων σε μετα-επίπεδο σχηματίζεται με βάση τις προβλέψεις του ταξινομητή του βασικού επιπέδου, στα οποία εφαρμόζεται τελικά ο ταξινομητής του μετα-επιπέδου.

2.3.3 Εναλλακτικές μέθοδοι συνδυασμού ταξινομητών

Εκτός από τεχνικές *ψηφοφορίας* και *συσσωρευμένης γενίκευσης*, υπάρχει άλλη μια κύρια τεχνική συνδυασμού που ονομάζεται *διαδοχική γενίκευση (cascade generalization -cascading,* [54]*)*. Κατά τη μέθοδο αυτή οι ταξινομητές εκπαιδεύονται ακολουθιακά ενώ δεν υφίσταται η έννοια του ενός ταξινομητή σε μετα-επίπεδο όπως στη συσσωρευμένη

γενίκευση. Κάθε ταξινομητής όταν εφαρμόζεται σε ένα σύνολο διανυσμάτων χαρακτηριστικών, προσθέτει τις προβλέψεις του (πιθανοτική κατανομή σε όλες τις κλάσεις) στο ίδιο σύνολο επιστρέφοντας έτσι ένα νέο σύνολο διανυσμάτων το οποίο θα χρησιμοποιηθεί για την εκπαίδευση ενός επόμενου ταξινομητή.

Η εκπαίδευση δηλαδή στη διαδοχική γενίκευση είναι ακολουθιακή, σε αντίθεση με τη συσσωρευμένη γενίκευση όπου η εκπαίδευση γίνεται παράλληλα. Στη διαδοχική γενίκευση κάθε ταξινομητής έχει πρόσβαση στα χαρακτηριστικά των διανυσμάτων του βασικού επιπέδου, μαζί με τα χαρακτηριστικά που αντιστοιχούν στις προβλέψεις του προηγούμενου ταξινομητή, σε αντίθεση με τη συσσωρευμένη γενίκευση όπου ο ταξινομητής του μετα-επιπέδου έχει πρόσβαση μόνο στις προβλέψεις των ταξινομητών του βασικού επιπέδου. Βέβαια θα μπορούσαν να χρησιμοποιηθούν τα χαρακτηριστικά του βασικού επιπέδου και στα διανύσματα σε μετα-επίπεδο στη συσσωρευμένη γενίκευση *(bi-level stacking)*, αλλά ο Schaffer [93] έχει αμφισβητήσει τη χρησιμότητα αυτής της προσέγγισης. Επίσης οι Gama & Brazdil [54] αναφέρουν ότι η διαδοχική γενίκευση δεν απαιτεί διασταυρωμένη επικύρωση στα δεδομένα εκπαίδευσης, όπως γίνεται στη συσσωρευμένη γενίκευση για τη δημιουργία των δεδομένων του μετα-επιπέδου. Επομένως η διαδοχική γενίκευση θα είναι ταχύτερη. Κάτι τέτοιο δεν είναι απόλυτα τεκμηριωμένο, διότι και στη διαδοχική γενίκευση απαιτείται η δημιουργία συνόλου δεδομένων εκπαίδευσης σε μετα-επίπεδο και η διαδικασία διασταυρωμένης επικύρωσης ενδείκνυται για το σκοπό αυτό. Η αξιολόγηση της διαδοχικής γενίκευσης από τους Gama & Brazdil [54] έδειξε μεν καλύτερα αποτελέσματα από τη συσσωρευμένη γενίκευση, περιορίζεται όμως στη χρήση ταξινομητών που βασίζονται στα δέντρα απόφασης και στη θεωρία του Bayes.

Το σημαντικότερο μειονέκτημα της διαδοχικής γενίκευσης έχει να κάνει με τη σειρά με την οποία πρέπει να εκπαιδευτούν οι ταξινομητές. Αλλάζοντας την ακολουθία των ταξινομητών οδηγεί σε σημαντικές αλλαγές και στα αποτελέσματα αξιολόγησης. Επομένως, η βέλτιστη ακολουθία εκπαίδευσης των ταξινομητών δεν είναι γνωστή εκ των προτέρων, ενώ δοκιμάζοντας όλους τους δυνατούς συνδυασμούς οδηγεί σε τεράστια αύξηση του υπολογιστικού κόστους. Από την άλλη πλευρά, δεν υφίσταται τέτοιο πρόβλημα στη συσσωρευμένη γενίκευση, όπου οι ταξινομητές εκπαιδεύονται *παράλληλα*. Επιπλέον, η διαδοχική γενίκευση αυξάνει τη *διαστατικότητα (dimensionality)* του συνόλου εκπαίδευσης, αφού σε κάθε βήμα της ακολουθιακής εκπαίδευσης προστίθενται καινούρια χαρακτηριστικά στα διανύσματα. Αντίθετα, η διαστατικότητα των διανυσμάτων σε μετα-επίπεδο στη συσσωρευμένη γενίκευση, δεν εξαρτάται από τη

διαστατικότητα των διανυσμάτων του βασικού επιπέδου αλλά από τον αριθμό των ταξινομητών του βασικού επιπέδου και τον αριθμό των σχετικών κλάσεων, αφού κάθε ταξινομητής επιστρέφει ένα διάνυσμα πιθανοτικών τιμών για όλες τις κλάσεις.

Μια μορφή διαδοχικής γενίκευσης πραγματοποιείται και στη μέθοδο της *ενδυνάμωσης (boosting)*, καθώς όπως υπενθυμίζεται και στην παράγραφο 2.2.2, μια ακολουθία ταξινομητών παράγεται (χρησιμοποιώντας τον ίδιο αλγόριθμο μάθησης) ακολουθιακά, με αλλαγή στα βάρη των διανυσμάτων εκπαίδευσης σε κάθε βήμα της ακολουθίας, σύμφωνα με τα λάθη των ταξινομητών που έχουν εκπαιδευτεί στα προηγούμενα βήματα. Ο Πίνακας 2.1 συνοψίζει τις βασικές μεθόδους συνδυασμού ταξινομητών που αναφέρθηκαν στις παραγράφους 2.2 και 2.3, μαζί με ενδεικτική βιβλιογραφία.

Πίνακας 2.1 Σύνοψη μεθόδων συνδυασμού ταξινομητών, μαζί με ενδεικτική βιβλιογραφία.

Μεθοδολογία συνδυασμού	*Ενδεικτική βιβλιογραφία*
Voting	[9, 13, 37, 51, 86, 101, 112]
Stacked generalization	[14, 19, 41, 78, 93, 98, 107, 110, 120]
Cascade generalization	[54]

2.4 Εξαγωγή πληροφορίας

Η *εξαγωγή πληροφορίας (information extraction)* είναι μια νέα περιοχή έρευνας που αναπτύσσεται ταχύτατα τα τελευταία χρόνια και τοποθετείται μεταξύ των περιοχών της *ανάκτησης πληροφορίας (information retrieval, IR)* και της *επεξεργασίας φυσικής γλώσσας (natural language processing, NLP)*. Αντίθετα με την ανάκτηση πληροφορίας, όπου το πρόβλημα είναι να εντοπιστούν (κατηγοριοποιηθούν) τα σχετικά έγγραφα κειμένων από μια ευρύτερη συλλογή, στη εξαγωγή πληροφορίας το πρόβλημα είναι να εντοπιστεί και να εξαχθεί η επιθυμητή πληροφορία μέσα σε ένα έγγραφο κειμένου. Αντίθετα επίσης με την περιοχή της επεξεργασίας φυσικής γλώσσας, όπου στην πιο απαιτητική της –και υπολογιστικά ακριβότερη- μορφή περιλαμβάνει την κατανόηση σε βάθος ενός εγγράφου κειμένου, η εξαγωγή πληροφορίας θεωρείται ως ένα είδος “φθηνότερης” –και σε υπολογιστικό κόστος- κατανόησης ενός εγγράφου, όσο απαιτείται για να συμπληρωθεί το σχεδιότυπο ενός εγγράφου, ή μια παραδοσιακή βάση δεδομένων, με την επιθυμητή πληροφορία.

Αρχικά δίνεται ένας ορισμός του προβλήματος της εξαγωγής πληροφορίας, στη συνέχεια περιγράφεται συνοπτικά η σχετική βιβλιογραφία, ενώ στο τέλος εξηγείται γιατί η εξαγωγή πληροφορίας θεωρείται ένα πρόβλημα εξόρυξης γνώσης από δεδομένα.

2.4.1 Ορισμός

Για τον ορισμό της εξαγωγής πληροφορίας προ-απαιτούνται οι εξής ορισμοί: Έστω $\{f^1 ... f^Q\}$ ένα σύνολο Q *πεδίων (fields)* εξαγωγής πληροφορίας σχετικών για μια συγκεκριμένη περιοχή ενδιαφέροντος και d ένα κείμενο επισημειωμένο από τον *ειδικό της θεματικής περιοχής (domain expert)* με *παραδείγματα (instances)* των σχετικών πεδίων. Ένα παράδειγμα πεδίου είναι ένα ζευγάρι $< t(s,e), f >$, όπου το $t(s,e)$ είναι ένα τμήμα κειμένου, με τους δείκτες *s* και *e* να είναι τα όρια έναρξης και τέλους αντίστοιχα του συγκεκριμένου τμήματος στον πίνακα των *λεκτικών μονάδων (tokens)* του κειμένου και $f \in \{f^1 ... f^Q\}$ είναι το πεδίο που σχετίζεται με το τμήμα κειμένου. Ένα *όριο (boundary)* ορίζεται ως το εικονικό διάστημα μεταξύ δύο γειτονικών λεκτικών μονάδων. Έστω T ένα *σχεδιότυπο (template)* το οποίο συμπληρώνεται με ζευγάρια $< t(s,e), f >$. Ένα πεδίο είναι τυπικά μια *θέση-στόχο (target-slot)* στο σχεδιότυπο T, ενώ ένα τμήμα κειμένου $t(s,e)$ είναι ένας *γεμιστήρας-θέσης (slot-filler)*. Ένα πεδίο μπορεί να έχει πολλαπλές εμφανίσεις ή καμία εμφάνιση μέσα σε ένα κείμενο.

Ο Πίνακας 2.2(α) δείχνει ένα τμήμα μιας ιστοσελίδας η οποία περιγράφει ένα προϊόν φορητού ηλεκτρονικού υπολογιστή, όπου τα σχετικά (επισημειωμένα) τμήματα κειμένου είναι τονισμένα με έντονη γραφή. Ο Πίνακας 2.2(β) δείχνει το σχεδιότυπο που αντιστοιχεί στο κείμενο του Πίνακα 2.2(α) και το οποίο έχει συμπληρωθεί χειρονακτικά από τον ειδικό της θεματικής περιοχής.

Πίνακας 2.2 (α) Τμήμα μιας σελίδας κειμένου του παγκοσμίου ιστού που περιγράφει προϊόντα φορητών υπολογιστών (β) το χειρονακτικά συμπληρωμένο σχεδιότυπο για τη σελίδα από τον ειδικό της θεματικής περιοχής.

…**TransPort ZX**
 <font size="1"> **15"** XGA **TFT** Display
 Intel Pentium III 600 MHZ 256k Mobile processor
  256 MB SDRAM up to 1GB
  40 GB hard drive (removable)
 …

(α)

Σχεδιότυπο T			Σύντομη περιγραφή για το πεδίο f
$t(s,e)$	s, e	Πεδίο f	
TransPort ZX	47, 49	model	Όνομα μοντέλου του φορητού
15"	56, 58	screenSiz	Μέγεθος της οθόνης του φορητού
TFT	59, 60	screenTyp	Τύπος της οθόνης του φορητού
IntelPentium	63, 67	procName	Όνομα του επεξεργαστή του φορητού
600 MHZ	67, 69	procSpee	Ταχύτητα του επεξεργαστή του φορητού
256 MB	76, 78	ram	Χωρητικότητα μνήμης RAM του φορητού
40 GB	86, 88	HDcapacit	Χωρητικότητα σκληρού δίσκου του

(β)

Το πρόβλημα της εξαγωγής πληροφορίας μπορεί να οριστεί ως εξής: *δοθέντος ενός κειμένου d, βρες όλα τα δυνατά παραδείγματα για κάθε σχετικό πεδίο μέσα στο d, και συμπλήρωσε το σχεδιότυπο που αντιστοιχεί στο d.* Ο ορισμός αυτός ορίζει ότι κάθε πρόβλημα μάθησης ενός σχετικού πεδίου αντιμετωπίζεται ξεχωριστά και μοντελοποιείται ως ένα δυαδικό πρόβλημα μάθησης: δοθέντος ενός αλγορίθμου μηχανικής μάθησης σχεδιασμένου για εξαγωγή πληροφορίας, τότε για κάθε σχετικό πεδίο $f \in \{f^1 ... f^Q\}$ μια *έννοια στόχο (target-concept)* μαθαίνεται για την αναγνώριση σχετικών παραδειγμάτων $< t(s,e), f >$ μέσα σε κείμενα. Κατά τη διαδικασία επαλήθευσης κάθε *έννοια στόχος* εφαρμόζεται χωριστά σε ένα κείμενο d, και το σχεδιότυπο T για το d συμπληρώνεται με τα αναγνωρισμένα παραδείγματα $< t(s,e), f >$ μέσα στο κείμενο.

Μια επέκταση του προβλήματος της εξαγωγής πληροφορίας είναι η μελέτη αλληλεπιδράσεων ανάμεσα στα σχετικά πεδία, που αναφέρεται και ως *εξαγωγή πολλαπλών θέσεων (multi-slot extraction,* [106]*)*. Σε αυτή τη διατριβή χειριζόμαστε την απλούστερη περίπτωση της *εξαγωγής μονής θέσης (single-slot extraction)*, η οποία καλύπτει ένα μεγάλο εύρος προβλημάτων εξαγωγής πληροφορίας και αποτέλεσε κίνητρο για την ανάπτυξη πληθώρας αλγορίθμων, όπως για παράδειγμα οι αλγόριθμοι που περιγράφονται στις εργασίες [16, 24, 48, 49].

Πρέπει να σημειωθεί ότι σε μερικές περιπτώσεις αλγορίθμων (για παράδειγμα [24]), η πληροφορία που αφορά τα όρια έναρξης και τέλους *s* και *e* αντίστοιχα κάθε τμήματος κειμένου $t(s,e)$ αναπαρίσταται με έμμεσο τρόπο μέσα σε ένα κείμενο, μέσω της παρουσίας κατάλληλων XML ετικετών, για παράδειγμα $< f >$ και $</ f >$ ετικετών για κάθε σχετικό πεδίο $f \in \{f^1 ... f^Q\}$. Στην περίπτωση αυτή, ένα σχεδιότυπο σαν κι αυτό του Πίνακα 2.2(β) μπορεί εύκολα να κατασκευαστεί.

2.4.2 Διεθνής επισκόπηση

Προγενέστερα της εφαρμογής της στον παγκόσμιο ιστό προς την κατεύθυνση της αντιμετώπισης του προβλήματος της υπερ-πληροφόρησης, η εξαγωγή πληροφορίας βρήκε πεδίο εφαρμογής σε ελεύθερο κείμενο με ειδησεογραφικό περιεχόμενο.

Εξαγωγή πληροφορίας από ελεύθερο κείμενο – Τα συνέδρια MUC

Το πρόβλημα της εξαγωγής πληροφορίας ήταν στο επίκεντρο των συνεδρίων *MUC* (Message Understanding Conferences, π.χ. [35, 36]). Οι θεματικές περιοχές στις οποίες αξιολογήθηκαν τα συστήματα που συμμετείχαν αφορούσαν συλλογές ελεύθερου (αδόμητου) κειμένου. Παράδειγμα αποτελεί μια συλλογή άρθρων εφημερίδων που

περιγράφουν μια ή περισσότερες τρομοκρατικές πράξεις σε χώρες της Λατινικής Αμερικής. Ένα άλλο παράδειγμα είναι μια συλλογή ειδησεογραφικών άρθρων με εξαγορές και συγχωνεύσεις εταιρειών, όπου θα πρέπει να αναγνωριστούν όλα τα τμήματα κειμένου που αντιστοιχούν σε ονόματα εταιρειών, καθώς και να προσδιοριστεί ο ρόλος της κάθε εταιρείας. Για παράδειγμα ποια είναι η εταιρεία αγοραστής, ποια είναι η αγοραζόμενη εταιρεία, σε ποια νέα εταιρεία συγχωνεύονται δύο άλλες, κλπ.

Στην πιο απαιτητική της μορφή, η εξαγωγή πληροφορίας περιλαμβάνει μια ακολουθία από στάδια, όπως η αναγνώριση και επισημείωση λεκτικών μονάδων, η συντακτική ανάλυση, η αναγνώριση ονομάτων οντοτήτων, ο προσδιορισμός ρόλων κ.α. [17]. Στα περισσότερα από αυτά τα στάδια έχουν εφαρμοστεί τεχνικές μηχανικής μάθησης με κύριο στόχο τη μείωση του χρόνου ανάπτυξης ενός πλήρους συστήματος εξαγωγής πληροφορίας και την εύκολη προσαρμογή του σε νέες θεματικές περιοχές. Ενδεικτικά αναφέρονται οι εργασίες [15, 57, 74] όπου προτείνεται η χρήση μηχανικής μάθησης για την επισημείωση λεκτικών μονάδων με μέρη του λόγου. Επίσης και το στάδιο της αναγνώρισης ονομάτων οντοτήτων έχει αποτελέσει πεδίο ευρείας εφαρμογής τεχνικών μηχανικής μάθησης (ενδεικτικά, [1, 10, 42]). Η ύπαρξη πολλαπλών σταδίων σε ένα πρόβλημα εξαγωγής πληροφορίας στην πιο απαιτητική του μορφή, συνεπάγεται και την ύπαρξη πιο πολύπλοκων σχεδιοτύπων από αυτό που εικονίζεται στον Πίνακα 2.2(β).

Εξαγωγή πληροφορίας από τον ιστό - Wrappers

Από την άλλη πλευρά, η ταχεία εξάπλωση του παγκοσμίου ιστού έχει εντείνει την ανάγκη ανάπτυξης συστημάτων τα οποία θα βοηθούν το χρήστη να χειριστεί τον τεράστιο όγκο δεδομένων κειμένου που είναι διαθέσιμος στον ιστό. Συστήματα τα οποία εξάγουν πληροφορία από τον ιστό θα πρέπει γενικά να πληρούν προδιαγραφές χαμηλού κόστους και μεγάλης ευελιξίας στην ανάπτυξη και προσαρμογή τους σε νέες θεματικές περιοχές. Τα συστήματα που αναπτύχθηκαν στα πλαίσια των συνεδρίων MUC αποτυγχάνουν γενικά να ανταποκριθούν στις προδιαγραφές αυτές, συν το γεγονός ότι η γλωσσική επεξεργασία που εφαρμόζεται για θεματικές περιοχές ελευθέρου κειμένου δεν εκμεταλλεύεται την εξτρα-γλωσσική πληροφορία που είναι διαθέσιμη στις σελίδες του παγκοσμίου ιστού (HTML/XML ετικέτες). Για τους λόγους αυτούς, τα MUC συστήματα δεν έχουν βρει μεγάλη ανταπόκριση στο χώρο της εξαγωγής πληροφορίας από τον παγκόσμιο ιστό.

Παρόλο που τα διακεκριμένα στάδια στην εξαγωγή πληροφορίας που αναφέρθηκαν προηγουμένως ότι υπάρχουν σε ένα τυπικό MUC σύστημα μπορούν να οριστούν και στην περίπτωση των εγγράφων του παγκοσμίου ιστού, πρακτικά δεν ισχύει κάτι τέτοιο.

Στον ιστό προσπαθούμε να εκμεταλλευτούμε περισσότερο τη δομή που υπάρχει σε ένα HTML ή XML έγγραφο για να επιλύσουμε ένα πρόβλημα εξαγωγής πληροφορίας. Σε αρκετά δομημένα ή ημι-δομημένα έγγραφα του ιστού τυπικά υπάρχει πολύ λίγο ελεύθερο (αδόμητο) κείμενο το οποίο περιέχει την επιθυμητή πληροφορία για την εξαγωγή. Ως εκ τούτου απαιτείται σχεδόν καθόλου ή πολύ λιγότερη γλωσσική πληροφορία για την πραγματοποίηση της εξαγωγής από ότι μπορεί να εκμεταλλευτεί ένα τυπικό σύστημα MUC. Επομένως τα σχεδιότυπα για εξαγωγή πληροφορίας από τον ιστό δεν υιοθετούν την πολυπλοκότητα των αντίστοιχων για τα MUC και τυπικά έχουν τη μορφή του Πίνακα 2.2(β).

Ως αποτέλεσμα, λιγότερο γλωσσικά διεξοδικές προσεγγίσεις αναπτύχθηκαν για την εξαγωγή πληροφορίας από τον παγκόσμιο ιστό, οι οποίες βασίζονται στη χρήση ειδικών προγραμμάτων εξαγωγής *(wrappers)*, τα οποία είναι σύνολα κανόνων που αναγνωρίζουν με μεγάλη ακρίβεια συγκεκριμένα τμήματα κειμένου σε ένα αρχείο. Η χειρονακτική κατασκευή των *wrappers* [21] αποδείχτηκε ως μια ιδιαίτερη επίπονη, χρονοβόρα και γεμάτη λάθη διαδικασία, η οποία απαιτεί μεγάλο βαθμό εμπειρογνωμοσύνης. Τεχνικές μηχανικής μάθησης έχουν χρησιμοποιηθεί για την εκμάθηση *wrappers* για εξαγωγή πληροφορίας, είτε με χρήση επιβλεπόμενων τεχνικών μάθησης (ενδεικτικά [29, 58, 69, 82]) ή με χρήση μη-επιβλεπόμενων τεχνικών μάθησης (ενδεικτικά, [20, 32]). Επίσημα, ένας *wrapper* ορίζεται [21] ως μια συνάρτηση από μια σελίδα κειμένου σε ένα σύνολο δομημένων πλειάδων που περιέχουν την επιθυμητή πληροφορία. Το σύνολο αυτό των πλειάδων αντιστοιχεί ουσιαστικά στις καταχωρήσεις σε ένα σχεδιότυπο, όπως αυτό του Πίνακα 2.2(β).

Οι *wrappers* προορίζονται για την εξαγωγή πληροφορίας από ισχυρά δομημένες σελίδες του ιστού, όπως τηλεφωνικοί κατάλογοι και κατάλογοι προϊόντων. Εκτός τούτου, οι σελίδες στις οποίες απευθύνονται οι *wrappers* πρέπει να μοιράζονται την ίδια δομή. Επομένως μια βιβλιοθήκη από *wrappers* μπορούν να κατασκευαστούν ανάλογα με το είδος και τη δομή της πληροφορίας που θέλουμε να εξάγουμε. Για παράδειγμα, μπορεί να κατασκευαστεί ένας *wrapper* που να εξάγει αριθμούς τηλεφώνων από αντίστοιχη σελίδα καταλόγου του δικτυακού τόπου *yahoo*[1], κι ένας διαφορετικός *wrapper* για την εξαγωγή τίτλων μουσικών άλμπουμ από σελίδες του ίδιου τόπου.

Οι *wrappers* αποτελούν αναπόσπαστο τμήμα εφαρμογών *μεσολαβητών πληροφορίας (information mediators)* οι οποίες συλλέγουν και επεξεργάζονται δεδομένα από διαφορετικές πηγές δεδομένων στον ιστό και παρουσιάζουν το αποτέλεσμα στον τελικό

[1] Yahoo, http://www.yahoo.com

χρήστη (ενδεικτικά [21, 40, 64, 77]). Χαρακτηριστικό παράδειγμα αποτελούν εφαρμογές *πρακτόρων σύγκρισης αγοράς (shopping comparison agents)* οι οποίες αναλαμβάνουν να επιστρέψουν στον τελικό χρήστη συγκριτικά χαρακτηριστικά, από διαφορετικούς *τόπους πωλητών (vendor sites)* ενός συγκεκριμένου προϊόντος που απαιτεί ο χρήστης. Μια γνωστή εφαρμογή αυτού του είδους υπήρξε το *ShopBot* [40].

Οι Muslea et al. [82] παρουσίασαν τον αλγόριθμο STALKER για την εκμάθηση *wrappers*. Ο STALKER είναι ένας απλός αλγόριθμος *ακολουθιακής επικάλυψης (sequential covering)* για την εκμάθηση κανόνων εξαγωγής πληροφορίας από δομημένα κείμενα του ιστού. Το ιδιαίτερο χαρακτηριστικό του αλγορίθμου αυτού είναι η χρήση ενός ειδικού φορμαλισμού με διακριτικό τίτλο *Embedded Catalog (EC)*, για τη μοντελοποίηση σελίδων του παγκοσμίου ιστού με κοινή δομή στο περιεχόμενό τους. Χρησιμοποιώντας το φορμαλισμό αυτό ως οδηγό, κοινοί *wrappers* προκύπτουν από τα δεδομένα εκπαίδευσης και χρησιμοποιούνται κατά την επαλήθευση για την εξαγωγή πληροφορίας. Το σημαντικότερο πλεονέκτημα της προσέγγισης αυτής είναι η δυνατότητα εξαγωγής πληροφορίας από σελίδες με ιεραρχικό περιεχόμενο (εξαιτίας του φορμαλισμού *EC* που περιγράφει την ιεραρχική αυτή δομή), στις οποίες προηγούμενες προσεγγίσεις για την κατασκευή *wrappers* αποτυγχάνουν [58, 69]. Το μειονέκτημα όμως είναι ότι πρέπει διαφορετικοί φορμαλισμοί *EC* πρέπει να κατασκευάζονται χειρονακτικά για σελίδες που δεν έχουν την ίδια δομή.

Η τελευταία παρατήρηση αφορά ένα σημαντικό μειονέκτημα της χρήσης των *wrappers* για εξαγωγή πληροφορίας. Στον παγκόσμιο ιστό είναι πολύ συχνό το φαινόμενο αρκετά δομημένες σελίδες που ανήκουν σε μια θεματική περιοχή να έχουν διαφορετική δομή, ανάλογα με τον δικτυακό τόπο στον οποίο φιλοξενούνται. Πάλι διαφορετικοί *wrappers* θα πρέπει να κατασκευαστούν στην περίπτωση αυτή για κάθε διαφορετική δομή. Επίσης είναι αρκετά συχνό το φαινόμενο ένας δικτυακός τόπος να αλλάζει σε τακτά χρονικά διαστήματα τη δομή των σελίδων του, καθιστώντας με τον τρόπο αυτό άχρηστους πλέον τους ήδη υπάρχοντες *wrappers* που βασίζονται σε μια συγκεκριμένη δομή για την εξαγωγή πληροφορίας. Στα [70, 71] συζητείται το θέμα της *συντήρησης των wrappers (wrapper maintenance)*. Το ζητούμενο στην περίπτωση αυτή, είναι η δυνατότητα εντοπισμού των αλλαγών στη δομή των σελίδων από τις οποίες εξάγεται η επιθυμητή πληροφορία και η αυτόματη αναπροσαρμογή των *wrappers* στη νέα δομή ώστε να συνεχιστεί ομαλά η εξαγωγή. Εάν κάτι τέτοιο δεν είναι δυνατό, ο χρήστης θα κληθεί να προσαρμόσει χειρονακτικά τους *wrappers* στη νέα δομή.

Στην εργασία [28] παρουσιάζεται ένα σύστημα εκμάθησης ευριστικών για την εξαγωγή πληροφορίας από δομημένες σελίδες που βρίσκονται σε διαφορετικούς δικτυακούς τόπους, τα οποία ευριστικά δεν θα εξαρτώνται από τη δομή της εκάστοτε σελίδας. Η είσοδος, όμως, του συστήματος αυτού, πρέπει να είναι ένα σύνολο ήδη υπαρχόντων *wrappers*, ταιριασμένων με τις σελίδες στις οποίες εφαρμόζονται. Στην εργασία [29] περιγράφεται ένα κομμάτι ενός εμπορικά εκμεταλλεύσιμου συστήματος εξαγωγής πληροφορίας από δομημένες σελίδες του ιστού που συλλέγονται από διαφορετικούς δικτυακούς τόπους. Το συγκεκριμένο σύστημα επικεντρώνεται στην εξαγωγή πληροφορίας που βρίσκεται σε στοιχεία πινάκων και λιστών. Εάν όμως κατά την εξαγωγή παρουσιαστεί ένας πίνακας ή μια λίστα με δομή αρκετά διαφορετική από τις δομές των αντίστοιχων στοιχείων που έχουν συναντηθεί κατά την εκπαίδευση, η εξαγωγή θα αποτύχει. Το πλεονέκτημα είναι ότι το σύστημα μπορεί να προσαρμοστεί, χειρονακτικά αλλά εύκολα, ώστε στα νέα προβλήματα που προκύπτουν κατά την εξαγωγή πληροφορίας από τον παγκόσμιο ιστό.

Εξαγωγή πληροφορίας από τον ιστό – Εργαλεία βασισμένα στην HTML και γλώσσες συγγραφής *wrappers*

Υπάρχει πληθώρα εργαλείων στη διεθνή βιβλιογραφία τα οποία βασίζονται σε δομικά γνωρίσματα της γλώσσας HTML για την εξαγωγή πληροφορίας από ισχυρά δομημένες σελίδες του ιστού. Πριν την εξαγωγή, η σελίδα μετασχηματίζεται σε ένα δέντρο το οποίο αναπαριστά την ιεραρχία των HTML ετικετών. Ο χρήστης, στη συνέχεια, με τη βοήθεια ενός ειδικού γραφικού περιβάλλοντος, είναι υπεύθυνος για τη σύνταξη των κανόνων εξαγωγής πληροφορίας (προγράμματα *wrappers*) με βάση τη δενδρική αυτή δομή και χρησιμοποιώντας μια ειδική γλώσσα. Παραδείγματα τέτοιων εργαλείων είναι τα *W4F* (World Wide Web Wrapper Factory, [92]), *XWRAP* [76] και *NoDoSE* [98].

Σημαντική ερευνητική δραστηριότητα υπάρχει και στις ειδικές γλώσσες για τη συγγραφή προγραμμάτων *wrappers* (ενδεικτικά, [7, 21, 31]). Για παράδειγμα, η γλώσσα *WebOQL* στην τελευταία εργασία πραγματοποιεί ερωτήματα τύπου SQL σε σελίδες του ιστού για τον εντοπισμό της σχετικής πληροφορίας. Συγκεκριμένα, και όπως αναφέρθηκε προηγουμένως, μια σελίδα του ιστού μετασχηματίζεται σε ένα δέντρο το οποίο αναπαριστά την HTML δενδρική δομή της σελίδας. Στη συνέχεια, χρησιμοποιώντας τη γλώσσα αυτή, ο χρήστης έχει τη δυνατότητα να συντάξει κανόνες εντοπισμού της σχετικής πληροφορίας πάνω στη δενδρική αυτή δομή.

Ένα σημαντικό μειονέκτημα των παραπάνω προσεγγίσεων είναι ο μεγάλος βαθμός συμμετοχής του χρήστη στη διαδικασία σύνταξης των κανόνων εξαγωγής, όπου

διαφορετικοί *wrappers* απαιτούνται για διαφορετικές δομές σελίδων. Το σύστημα XWRAP χρησιμοποιεί για το σκοπό ένα σύνολο ευριστικών για να χειριστεί διαφορετικές δομές σε μια σελίδα του ιστού (πίνακες, λίστες κ.α.).

Εξαγωγή πληροφορίας από τον ιστό – Χρήση οντολογιών

Η εξαγωγή πληροφορίας από σελίδες του ιστού μπορεί επίσης να πραγματοποιηθεί με τη χρήση *οντολογιών*. Για παράδειγμα, το σύστημα *BYU* [42] καθώς και το σύστημα που περιγράφεται στην εργασία [34] χρησιμοποιούν οντολογίες οι οποίες διαθέτουν χειρονακτικά κατασκευασμένους κανόνες για τον εντοπισμό της σχετικής πληροφορίας σε ένα έγγραφο του ιστού. Ένα πλεονέκτημα των προσεγγίσεων αυτών είναι ότι η διαδικασία εξαγωγής είναι αυτοματοποιημένη, ιδιαίτερα εάν η οντολογία είναι αρκετά αντιπροσωπευτική της θεματικής περιοχής που περιγράφει. Ένα άλλο πλεονέκτημα επίσης είναι ότι η εξαγωγή πληροφορίας μπορεί να πραγματοποιηθεί από σελίδες με διαφορετική δομή στο περιεχόμενό τους. Από την άλλη πλευρά, ένα μειονέκτημα είναι και πάλι η μεγάλη συμμετοχή του χρήστη, οποίος θα πρέπει να είναι εξειδικευμένος στην κατασκευή οντολογιών.

Η χρήση τεχνικών μηχανικής μάθησης μπορεί να λειτουργήσει συμπληρωματικά στην κατασκευή μιας οντολογίας. Για παράδειγμα, οι χειρονακτικά κατασκευασμένοι κανόνες της οντολογίας για τον εντοπισμό της σχετικής πληροφορίας από ένα έγγραφο του ιστού, μπορούν να αντικατασταθούν με κανόνες που μαθαίνονται με χρήση μηχανικής μάθησης, αντισταθμίζονται το κόστος για την κατασκευή της οντολογίας.

Εξαγωγή πληροφορίας από τον ιστό – Προσαρμοστικά συστήματα

Ένα σημαντικό μειονέκτημα των προγραμμάτων *wrappers* είναι ότι αποτυγχάνουν σε προβλήματα εξαγωγής πληροφορίας από λιγότερο δομημένο κείμενο ή αδόμητο (ελεύθερο) κείμενο, που είναι επίσης αρκετά συχνό φαινόμενο στον παγκόσμιο ιστό. Επίσης είναι αρκετά συχνό το φαινόμενο ύπαρξης σελίδων με μικτή δομή, δηλαδή σελίδων που περιέχουν και δομημένο και σχεδόν αδόμητο κείμενο. Οι ερευνητικές εργασίες [28, 29] αντιμετωπίζουν το πρόβλημα της εξαγωγής πληροφορίας από ισχυρά δομημένες σελίδες με μεταβαλλόμενη δομή.

Από τη άλλη πλευρά, η εκπαίδευση συστημάτων τα οποία θα αναγνωρίζουν την επιθυμητή πληροφορία τόσο από ισχυρά δομημένες σελίδες, όσο και από λιγότερο (ή καθόλου) δομημένες σελίδες αποτελεί ένα σημαντικό ζητούμενο από την ερευνητική κοινότητα της εξαγωγής πληροφορίας. Στην εργασία [73] γίνεται ένας διαχωρισμός των σελίδων του ιστού σε εκείνες που περιέχουν *ημι-δομημένα (semi-structured) δεδομένα*

και σε εκείνες που περιέχουν *ημι-δομημένο κείμενο*. Οι σελίδες της πρώτης κατηγορίας είναι ισχυρά δομημένες και περιέχουν αντικείμενα σχετικά με την εξαγωγή τα οποία είναι υποκρυπτόμενα στο εσωτερικό τους. Τέτοιες σελίδες τυπικά παράγονται δυναμικά από μια βάση δεδομένων, ως απάντηση στο ερώτημα ενός χρήστη (χαρακτηριστικό παράδειγμα είναι οι σελίδες απαντήσεων των μηχανών αναζήτησης). Οι *wrappers*, χειρίζονται καλύτερα σελίδες αυτού του τύπου. Αντίθετα, οι σελίδες της δεύτερης κατηγορίας περιέχουν και τμήματα λιγότερου δομημένου ή ελεύθερου κειμένου κι επομένως οι κοινοί *wrappers* δεν τις χειρίζονται ικανοποιητικά. Η εκμετάλλευση γλωσσικής πληροφορίας που είναι διαθέσιμη σε λιγότερο δομημένο κείμενο καθίσταται απαραίτητη για την εξαγωγή πληροφορίας από τέτοιες σελίδες.

Πρόσφατη έρευνα παροτρύνει την ανάπτυξη *προσαρμοστικών (adaptive)* συστημάτων εξαγωγής πληροφορίας [24, 26], τα οποία θα είναι σε θέση να χειριστούν κείμενα διαφορετικού τύπου δόμησης, από αρκετά δομημένο σε σχεδόν ελεύθερο κείμενο, συμπεριλαμβανομένων και μικτών τύπων. Για παράδειγμα, οι αλγόριθμοι που παρουσιάζονται στις εργασίες [16, 24, 106], μαθαίνουν κανόνες εξαγωγής οι οποίοι εκμεταλλεύονται πληροφορία που προέρχεται από επεξεργασία φυσικής γλώσσας και γι' αυτό μπορούν να εφαρμοστούν και σε λιγότερο δομημένο κείμενο, όπου τέτοιου είδους πληροφορία είναι διαθέσιμη. Ο αλγόριθμος BWI [49], μαθαίνει *wrappers* οι οποίοι εκμεταλλεύονται περιορισμένη γλωσσική πληροφορία, σε σχέση με τους αλγορίθμους στις προηγούμενες εργασίες. Όμως, μια μεθοδολογία που βασίζεται στην *ενδυνάμωση (boosting,* [51]*)*, ακολουθείται κατά τη διαδικασία μάθησης η οποία βελτιώνει την απόδοση στην εξαγωγή κι επιτρέπει την εφαρμογή του αλγορίθμου και σε λιγότερο δομημένο κείμενο. Επίσης, τα *Κρυφά Μαρκοβιανά μοντέλα* [91], μια ισχυρή στατιστική μέθοδος, έχει βρει μεγάλη εφαρμογή σε προβλήματα εξαγωγής πληροφορίας τόσο σε δομημένο όσο και σε ελεύθερο κείμενο [50, 100].

Ένα σημαντικό κίνητρο για την ανάπτυξη προσαρμοστικών συστημάτων εξαγωγής πληροφορίας είναι η χρήση τους για την εκπλήρωση του οράματος του *σημασιολογικού ιστού (semantic web)*. Η χρήση προτύπων σημασιολογικής επισημείωσης κειμένων, τα οποία πρότυπα βασίζονται στη γλώσσα XML, σε συνδυασμό με τη χρήση *οντολογιών*, αποσκοπούν στη δημιουργία *περιεχομένου κατανοητού από τη μηχανή (machine readable content)*, ώστε εφαρμογές *πρακτόρων (agents)* να κινούνται στο χώρο του ιστού, επεξεργαζόμενοι με διάφορους τρόπους το σημασιολογικά δομημένο περιεχόμενό του, ανάλογα με τις απαιτήσεις του χρήστη. Δυστυχώς το σενάριο αυτό απέχει πολύ από την πραγματικότητα κυρίως για ένα λόγο: η συντριπτική πλειοψηφία

των σελίδων του ιστού δεν είναι σημασιολογικά επισημειωμένες. Το κόστος για την χειρονακτική επισημείωση του τεράστιου αυτού όγκου δεδομένων είναι απαγορευτικό. Η εξαγωγή πληροφορίας, ως μια μέθοδος αυτόματου εντοπισμού σχετικών παραδειγμάτων πεδίων σε ηλεκτρονικά κείμενα μπορεί να παίξει σημαντικό ρόλο στην ανάπτυξη αυτόματων ή ημι-αυτόματων μεθόδων σημασιολογικής επισημείωσης κειμένων του ιστού [25], καθώς και στον εμπλουτισμό οντολογιών με καινούρια παραδείγματα πεδίων [113]. Η χρήση προσαρμοστικών συστημάτων εξαγωγής πληροφορίας για τους σκοπούς του σημασιολογικού ιστού καθίσταται απαιτητή, εξαιτίας της ταυτόχρονης ύπαρξης στον παγκόσμιο ιστό τόσο ισχυρά δομημένων, όσο και αδόμητων σελίδων κειμένου. Περισσότερη συζήτηση στο θέμα της χρήσης προσαρμοστικών συστημάτων εξαγωγής πληροφορίας για τους σκοπούς του σημασιολογικού ιστού υπάρχει στην εργασία [27].

Ενδιαφέρουσες επισκοπήσεις συστημάτων εξαγωγής πληροφορίας από δεδομένα του παγκοσμίου ιστού αποτελούν οι εργασίες [71, 73, 81]. Ο Πίνακας 2.3 συνοψίζει τα συστήματα εξαγωγής πληροφορίας που συζητήθηκαν στην παρούσα ενότητα, μαζί με ενδεικτική βιβλιογραφία στην οποία μπορεί να απευθυνθεί ο αναγνώστης.

Πίνακας 2.3 Σύνοψη συστημάτων εξαγωγής πληροφορίας, μαζί με ενδεικτική βιβλιογραφία.

Κατηγορία	*Ενδεικτική Βιβλιογραφία*
MUC	[35, 36]
Wrappers	IEPAD [20], [28, 29], RoadRunner [32], SoftMealy [58], WIEN [69], STALKER [82]
HTML-aware	NoDoSE [2], WebOQL [7], [21, 31], XWRAP [76], WWWF [92]
Ontology-based	BYU [42], [34, 113]
Adaptive	RAPIER [16], $(LP)^2$ [24], BWI [49], HMMs [50, 100], WHISK [106]

2.4.3 Η εξαγωγή πληροφορίας ως ένα πρόβλημα εξόρυξης γνώσης

Ο παγκόσμιος ιστός, με τον τεράστιο όγκο δεδομένων που διαθέτει και ο οποίος αυξάνεται με ιλιγγιώδη ρυθμό τα τελευταία χρόνια, αποτελεί ιδανικό πεδίο εφαρμογής τεχνικών *εξόρυξης γνώσης* από τα δεδομένα του. Ο όρος *εξόρυξη γνώσης από δεδομένα (data mining)* αναφέρεται γενικά στην εφαρμογή τεχνικών μηχανικής μάθησης για την ανακάλυψη νέας πολύτιμης γνώσης *(knowledge discovery)* από μεγάλο όγκο δεδομένων και τη διαχείριση του όγκου αυτού.

Τεχνικές εξόρυξης γνώσης από δεδομένα έχουν αποδείξει ότι μπορούν να εφαρμοστούν επιτυχώς σε μεγάλο όγκο δεδομένων κι έχουν χρησιμοποιηθεί με επιτυχία σε αρκετά πεδία, όπως στην ιατρική (ιατρική διάγνωση), στη μετεωρολογία (πρόβλεψη

καιρού), στο μάρκετινγκ, κ.α. Ένα τυπικό σύστημα εξόρυξης γνώσης από δεδομένα αποτελείται από τρία κύρια στάδια: α) Προεπεξεργασία δεδομένων, όπου γίνεται *καθαρισμός δεδομένων (data cleaning)* και η κωδικοποίησή τους ώστε να ταιριάζουν με τις προδιαγραφές των αλγορίθμων μάθησης που θα εφαρμοστούν, β) εφαρμογή των αλγορίθμων μηχανικής μάθησης, γ) ανάλυση των αποτελεσμάτων από την εφαρμογή των αλγορίθμων για την ανακάλυψη πολύτιμης γνώσης από τα δεδομένα.

Η *εξόρυξη γνώσης από δεδομένα του παγκοσμίου ιστού* καλείται *Web Mining*. Ο όρος αυτός χρησιμοποιήθηκε αρχικά από τον Etzioni [43] και αποτελεί το σταυροδρόμι συνάντησης αρκετών ερευνητικών περιοχών όπως είναι η τεχνητή νοημοσύνη, καθώς και υποκατηγορίες της τεχνητής νοημοσύνης όπως η μηχανική μάθηση, η επεξεργασία φυσικής γλώσσας, οι βάσεις δεδομένων η ανάκτηση πληροφορίας, η εξαγωγή πληροφορίας και η *μοντελοποίηση χρηστών (user modeling)*.

Οι Kosala & Blockeel [66] προτείνουν έναν διαχωρισμό του όρου Web Mining σε *Web Content Mining*, *Web Structure Mining,* και *Web Usage Mining*, ανάλογα με τον αν τα δεδομένα του παγκοσμίου ιστού αφορούν το περιεχόμενο, τη δομή και τη χρήση του αντίστοιχα. Τα δεδομένα περιεχομένου του ιστού είναι οι σελίδες (HTML, XML) που είναι δημοσιευμένες σε αυτόν. Τα δεδομένα δομής αφορούν τους *υπερ-συνδέσμους (hyperlinks)* ανάμεσα στις σελίδες του ιστού, οι οποίοι καθορίζουν τη δομή κάθε δικτυακού τόπου και πως συνδέεται ο τόπος αυτός με το υπόλοιπο μέρος του παγκοσμίου ιστού. Τα δεδομένα χρήσης στον παγκόσμιο ιστό αφορούν την κίνηση των χρηστών που επισκέπτονται έναν δικτυακό τόπο, όπως η προέλευσή τους, τις σελίδες προσπέλασης κάθε χρήστη, ημερομηνία και ώρα προσπέλασης κάθε σελίδας κ.α. Τα δεδομένα αυτά συγκεντρώνονται συνήθως από τους *εξυπηρετητές του παγκοσμίου ιστού (Web servers)* σε ειδικά αρχεία κειμένου. Η εργασία των Kosala & Blockeel [66] αποτελεί μια άριστη αρχική ενημέρωση στο χώρο της εξόρυξης γνώσης από τον ιστό.

Η εξαγωγή πληροφορίας από τις σελίδες του ιστού με χρήση τεχνικών μηχανικής μάθησης ανήκει στην κατηγορία της εξόρυξης γνώσης από δεδομένα περιεχομένου του ιστού *(Web content mining)*. Η χρήση τεχνικών ψηφοφορίας και συσσωρευμένης γενίκευσης για το συνδυασμό συστημάτων εξαγωγής πληροφορίας ανήκει επίσης στην κατηγορία της εξόρυξης γνώσης από δεδομένα περιεχομένου του ιστού. Στην μεν ψηφοφορία συνδυάζονται κανόνες που έχουν *εξορυχθεί* από την εφαρμογή διαφορετικών αλγορίθμων μηχανικής μάθησης που είναι σχεδιασμένοι για εξαγωγή πληροφορίας. Στη συσσωρευμένη γενίκευση επιχειρείται *εξόρυξη* νέων κανόνων εξαγωγής πληροφορίας σε μετα-επίπεδο με την εφαρμογή αλγορίθμων ταξινόμησης.

Θα πρέπει να τονίσουμε ότι η κλασσική προσέγγιση του *data mining*, υποθέτει ότι τα δεδομένα από τα οποία θα γίνει εξόρυξη της γνώσης βρίσκονται ήδη στη μορφή μιας σχεσιακής βάσης δεδομένων. Στον ιστό, όμως, τα δεδομένα περιεχομένου είναι στη μορφή ημι-δομημένων αρχείων κειμένου. Η εφαρμογή αλγορίθμων μηχανικής μάθησης στο πρόβλημα της εξαγωγής πληροφορίας αποσκοπεί στην *εξόρυξη* κανόνων από το περιεχόμενο των επισημειωμένων κειμένων για μια θεματική περιοχή, οι οποίοι κανόνες θα μπορούν να αναγνωρίζουν σχετικά τμήματα κειμένου σε καινούρια μη επισημειωμένα κείμενα κατά τη διαδικασία επαλήθευσης. Επομένως, το πρόβλημα της εξόρυξης γνώσης από δεδομένα περιεχομένου του ιστού, θεωρείται και ως ένα πρόβλημα *εξόρυξης γνώσης από κείμενο (text mining)*. Μια εναλλακτική προσέγγιση στην εξόρυξη γνώσης από δεδομένα περιεχομένου του ιστού (ή εξόρυξης γνώσης από κείμενο) είναι η χρήση μηχανικής μάθησης για την εξαγωγή πληροφορίας από κείμενα του ιστού και τη συμπλήρωση μιας βάσης δεδομένων, και η εκ νέου χρήση τεχνικών μηχανικής μάθησης για την ανακάλυψη πολύτιμης γνώσης από τα δεδομένα της βάσης. Στην περίπτωση αυτή η εξαγωγή πληροφορίας αναλαμβάνει ουσιαστικά το ρόλο της προ-επεξεργασίας δεδομένων, κατά την κλασσική προσέγγιση του *data mining*.

Στη διατριβή αυτή, θεωρούμε ότι η πολύτιμη γνώση που ανακαλύπτεται αφορά τους *κανόνες* που εντοπίζουν και εξάγουν την επιθυμητή πληροφορία από κείμενα του ιστού. Η θεώρηση αυτή είναι ιδιαίτερα σημαντική, διότι καλούμαστε να διαχειριστούμε έναν τεράστιο όγκο δεδομένων, όπως εκείνος του ιστού. Από την άλλη πλευρά, υπάρχουν διάφορα συστήματα [55, 84] τα οποία χρησιμοποιούν τεχνικές εξαγωγής πληροφορίας για τη συμπλήρωση της βάσης δεδομένων και στη συνέχεια εφαρμόζουν κλασσικές τεχνικές *data mining*.

2.5 Συνδυασμός συστημάτων για εξαγωγή πληροφορίας

Παρά το ολοένα και αυξανόμενο ενδιαφέρον στο συνδυασμό αλγορίθμων μηχανικής μάθησης και την εφαρμογή τους σε προβλήματα επεξεργασίας φυσικής γλώσσας, όπως η *επισημείωση μερών του λόγου (part-of-speech tagging,* [57]*)* η *αποσαφήνιση εννοιών-λέξεων (word-sense disambiguation,* [45*)*, το αντικείμενο αυτό δεν έχει απασχολήσει ιδιαίτερα τη διεθνή ερευνητική κοινότητα της εξαγωγής πληροφορίας.

2.5.1 Πολυστρατηγική μάθηση

Ουσιαστικά η μόνη εργασία που σχετίζεται με το συνδυασμό συστημάτων εξαγωγής πληροφορίας προέρχεται από τον Freitag [48], όπου το πρόβλημα της εξαγωγής μοντελοποιείται ως ένα πρόβλημα κοινής ταξινόμησης χρησιμοποιώντας τέσσερα

συστήματα σε βασικό επίπεδο, τα οποία στη συνέχεια συνδυάζονται χρησιμοποιώντας *πολυστρατηγική μάθηση (multistrategy learning)*.

Ο όρος *πολυστρατηγική μάθηση* αναφέρεται γενικά στο συνδυασμό πολλαπλών προσεγγίσεων μάθησης κάτω από έναν κοινό αλγόριθμο, και χρησιμοποιήθηκε κυρίως για το συνδυασμό *επαγωγικής* με *αναλυτική* μάθηση [79]. Ο Domingos [39] χρησιμοποίησε τον όρο *εμπειρική πολυστρατηγική μάθηση* (*empirical multistrategy learning*) για να διαχωρίσει την περίπτωση όπου όλοι οι αλγόριθμοι μάθησης είναι επαγωγικοί. Η βασική αντίληψη πίσω από τη χρήση εμπειρικής πολυστρατηγικής μάθησης είναι ο σχεδιασμός ενός νέου πολύπλοκου αλγορίθμου ο οποίος συνδυάζει με ευριστικό τρόπο ένα σύνολο επαγωγικών αλγορίθμων μάθησης, απαιτώντας όμως λεπτομέρειες υλοποίησης κάθε αλγορίθμου ξεχωριστά. Για παράδειγμα, ο αλγόριθμος RISE [39] ενοποιεί έναν αλγόριθμο *μάθησης βασισμένο στα παραδείγματα (instance-based learning)* κι έναν αλγόριθμο *μάθησης βασισμένο σε κανόνες (rule-based learning)* σε μια νέα υλοποίηση, σκοπεύοντας ταυτόχρονα να ξεπεράσει τους περιορισμούς και των δύο προσεγγίσεων.

Η συσσωρευμένη γενίκευση συνδυάζει ένα σύνολο πολλαπλών αλγορίθμων μάθησης κάτω από μια πιο χαλαρή δομή, μαθαίνοντας μόνο να συνδυάζει την έξοδο των εκπαιδευμένων ταξινομητών, δηλαδή χειρίζεται κάθε αλγόριθμο μάθησης ως *μαύρο κουτί (black box)*. Παρά την απλοϊκότητα της προσέγγισης αυτής, η συσσωρευμένη γενίκευση προσφέρει το αρκετά σημαντικό πλεονέκτημα της εύκολης επεκτασιμότητας σε περισσότερους ταξινομητές, τόσο σε βασικό όσο και σε μετα-επίπεδο, ανεξάρτητα από την εσωτερική δομή των αντίστοιχων αλγορίθμων. Από την άλλη μεριά, η διεξαγωγή ψηφοφορίας είναι ο απλούστερος τρόπος συνδυασμού ταξινομητών.

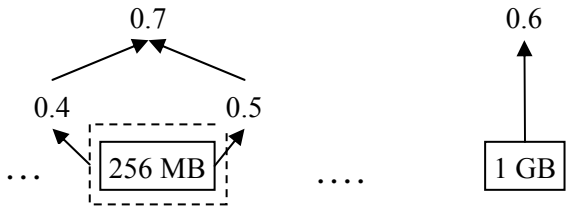
Ο Freitag [48] ακολούθησε το τελευταίο παράδειγμα για το συνδυασμό συστημάτων εξαγωγής πληροφορίας, το οποίο ονόμασε ως *πολυστρατηγική μάθηση για εξαγωγή πληροφορίας* και βασίζεται στη χρήση πιθανοτικών εκτιμήσεων στην έξοδο πολλαπλών συστημάτων εξαγωγής πληροφορίας. Πιο συγκεκριμένα, για κάθε σχετικό πεδίο $f \in \{f^1 \ldots f^Q\}$, οι *βαθμοί εμπιστοσύνης (confidence scores)* που παράγονται από τα συστήματα του βασικού επιπέδου μετασχηματίζονται σε πιθανότητες ορθότητας της ακρίβειας στις προβλέψεις κάθε συστήματος. Στην περίπτωση περισσοτέρων από μια προβλέψεις ενός πεδίου f για ένα τμήμα κειμένου $t(s,e)$, μια συνδυασμένη πιθανότητα υπολογίζεται για το παράδειγμα $< t(s,e), f >$ χρησιμοποιώντας την εξίσωση 2.1:

$$P^C = 1 - \prod_j (1 - p^j) \qquad (2.1)$$

όπου P^C είναι η συνδυασμένη πιθανότητα ότι το τμήμα κειμένου $t(s,e)$ ανήκει στο πεδίο f, και p^j είναι η πιθανότητα ότι κάποιο σύστημα εξαγωγής πληροφορίας E^j έχει προβλέψει το πεδίο f για το $t(s,e)$. Στην πραγματικότητα, το P^C αντιστοιχεί στην πιθανότητα ότι τουλάχιστον ένα σύστημα εξαγωγής πληροφορίας από αυτά που έχουν προβλέψει το πεδίο f για το $t(s,e)$ έχει προβλέψει σωστά, το οποίο ισούται με την πιθανότητα ότι δεν είναι όλες οι προβλέψεις για το f λανθασμένες.

Στη συνέχεια, ο περιορισμός του *ενός παραδείγματος ανά κείμενο (one per document, OPD)* εφαρμόζεται για μερικά από τα σχετικά πεδία της θεματικής περιοχής. Αυτό σημαίνει ότι μόνο ένα παράδειγμα για το συγκεκριμένο πεδίο για το οποίο ισχύει ο OPD περιορισμός μπορεί να εμφανίζεται σε μια σελίδα. Για παράδειγμα, μια σελίδα στη θεματική περιοχή των μαθημάτων της επιστήμης υπολογιστών, θα πρέπει να περιγράφει μόνο ένα μάθημα, και επομένως μόνο ένα παράδειγμα για το πεδίο “τίτλος μαθήματος” θα πρέπει να υπάρχει σε μια σελίδα. Εάν δύο ή περισσότερα παραδείγματα του πεδίου αυτού έχουν εντοπιστεί σε μια σελίδα, τότε εκείνο με τη μεγαλύτερη (συνδυασμένη) πιθανότητα επιλέγεται, ενώ όλα τα υπόλοιπα απορρίπτονται.

Το βασικό κίνητρο πίσω από αυτή τη στρατηγική συνδυασμού, όπως επεξηγήθηκε από τον Freitag [48], είναι ότι εάν οι προβλέψεις με το μεγαλύτερο σκορ για ένα πεδίο f από δύο ή περισσότερα συστήματα εξαγωγής πληροφορίας είναι όλες λανθασμένες, τότε μπορεί τα συστήματα αυτά να προβλέψουν, με ένα μικρότερο σκορ, παραδείγματα για το f τα οποία συμφωνούν στο σωστό τμήμα κειμένου $t(s,e)$. Επομένως, συνδυάζοντας τις προβλέψεις αυτές μπορούμε να καταλήξουμε στο σωστό παράδειγμα $< t(s,e), f >$. Στη συνέχεια, όλα τα υπόλοιπα παραδείγματα για το f απομακρύνονται. Το Σχήμα 2.3 δείχνει ένα χαρακτηριστικό παράδειγμα για το πώς δουλεύει η πολυστρατηγική μάθηση για εξαγωγή πληροφορίας.



Σχήμα 2.3 Συνδυάζοντας τις προβλέψεις δύο υποθετικών συστημάτων εξαγωγής για ένα παράδειγμα του πεδίου “ram”. Έστω ότι το σωστό παράδειγμα είναι $<"256\ MB", ram>$.

Κάθε διαφορετικός τύπος κουτιού στο Σχήμα 2.3 αντιπροσωπεύει ένα διαφορετικό υποθετικό σύστημα εξαγωγής πληροφορίας. Το σύστημα με το κουτί σε κανονικό τύπο

γραμμής αναγνωρίζει τα τμήματα κειμένου “256 MB” και “1 GB” ως παραδείγματα του πεδίου “ram”, ενώ το σύστημα με το κουτί στην διακεκομμένη γραμμή αναγνωρίζει μόνο το τμήμα κειμένου “256 MB” ως παράδειγμα του “ram”. Συνδυάζοντας τα δύο συστήματα, επιστρέφεται το παράδειγμα $<"256\ MB", ram>$ που έχει τη μεγαλύτερη συνδυασμένη πιθανότητα και από τα δύο συστήματα.

Ο OPD περιορισμός, παρόλο που μπορεί να φανεί χρήσιμος σε ορισμένες περιπτώσεις, είναι αρκετά περιοριστικός για το πρόβλημα της εξαγωγής και δεν ισχύει για όλα τα πεδία. Για παράδειγμα, μια ιστοσελίδα μπορεί να περιγράφει περισσότερα από ένα προϊόντα φορητών υπολογιστών, οπότε δεν είναι σωστό να διατηρηθεί μόνο ένα παράδειγμα όλων των πεδίων (“ram”, “μοντέλο”, “τιμή”, κ.λ.π.) και να απορριφθούν τα υπόλοιπα. Ο Freitag, μετά από προσωπική επικοινωνία, διευκρίνισε ότι εφάρμοσε τον OPD περιορισμό για όλα τα πεδία στις περιοχές που χρησιμοποίησε για αξιολόγηση, δίχως να υπάρχει σημαντική απώλεια πληροφορίας. Για παράδειγμα, μια σελίδα σπάνια περιγράφει περισσότερα από ένα μαθήματα επιστήμης υπολογιστών. Η προσέγγιση αυτή παραμένει παρόλα αυτά περιοριστική, αφού μια ιστοσελίδα συχνά περιγράφει περισσότερα από ένα προϊόντα φορητών ηλεκτρονικών υπολογιστών.

Η μετατροπή των βαθμών εμπιστοσύνης στις προβλέψεις ενός συστήματος εξαγωγής πληροφορίας σε πιθανοτικές εκτιμήσεις ορθότητας, πραγματοποιείται με τον έλεγχο της αξιοπιστίας του συστήματος σε ένα ξεχωριστό σύνολο δεδομένων ή με διαδικασία διασταυρωμένης επικύρωσης στα δεδομένα εκπαίδευσης. Κατόπιν χρησιμοποιείται μια μορφή μοντελοποίησης με χρήση *παλινδρόμησης (regression)* η οποία περιγράφεται με περισσότερες λεπτομέρειες στην εργασία [48].

Το κίνητρο για τη μετατροπή βαθμών εμπιστοσύνης σε πιθανότητες ορθότητας στις προβλέψεις ενός συστήματος εξαγωγής πληροφορίας εξηγείται ως εξής: κάθε σύστημα υπολογίζει με διαφορετικό τρόπο τους βαθμούς εμπιστοσύνης στις προβλέψεις του, ενώ και το εύρος τιμών των βαθμών εμπιστοσύνης είναι διαφορετικό για κάθε σύστημα. Απαιτείται λοιπόν μια διαδικασία κανονικοποίησης του εύρους τιμών για κάθε σύστημα στο διάστημα μεταξύ 0 και 1 ώστε να είναι δυνατή η διεξαγωγή άμεσων συγκρίσεων μεταξύ των συστημάτων. Επιπλέον οι βαθμοί εμπιστοσύνης δεν αντιστοιχούν πάντα σε πιθανότητες ορθότητας, δηλαδή δεν υπάρχει πάντα μια στενή συσχέτιση μεταξύ τους και μπορεί να οδηγήσει σε παραπλανητικά αποτελέσματα κατά το συνδυασμό συστημάτων εξαγωγής μέσω ψηφοφορίας. Για παράδειγμα, σε τμήματα κειμένου που έχουν αναγνωριστεί ως σχετικά είναι δυνατόν να εκχωρηθούν μεγαλύτερες τιμές εμπιστοσύνης από άλλα τμήματα κειμένου που έχουν ορθά αναγνωριστεί ως σχετικά. Η

διαδικασία της κανονικοποίησης θα πρέπει επομένως να περιλαμβάνει τον έλεγχο της αξιοπιστίας κάθε συστήματος σε κάποιο ξεχωριστό σώμα κειμένων ώστε οι νέες κανονικοποιημένες τιμές θα πρέπει να αντιστοιχούν σε πιθανότητες ορθότητας.

Οι Kauchak et al. [62] μελέτησαν τη συσχέτιση βαθμών εμπιστοσύνης σε πιθανότητες ορθότητες στις προβλέψεις ενός συστήματος εξαγωγής πληροφορίας που εκπαιδεύτηκε με χρήση του αλγορίθμου BWI. Το συμπέρασμά τους ήταν ότι η συσχέτιση αυτή είναι πιο αδύνατη για δύσκολα προβλήματα εξαγωγής πληροφορίας, για παράδειγμα από σελίδες με ελεύθερο κείμενο, οπότε η χρήση πιθανοτήτων ορθότητας είναι πιο σωστή.

Για το υπόλοιπο της διατριβής, θα χρησιμοποιηθεί ο όρος *πολυστρατηγική μάθηση*, αναφερόμενος στην μεθοδολογία συνδυασμού του Freitag [48].

2.5.2 Γιατί η χρήση ψηφοφορίας και συσσωρευμένης γενίκευσης;

Η επισκόπηση της διεθνούς βιβλιογραφίας οδηγεί στο συμπέρασμα ότι υπάρχει πολύ φτωχή ερευνητική δραστηριότητα μέχρι τώρα όσον αφορά το συνδυασμό συστημάτων εξαγωγής πληροφορίας, σε αντίθεση με το συνδυασμό κοινών ταξινομητών, όπου η δραστηριότητα είναι σημαντικά πλουσιότερη. Η εργασία [48] συνδυάζει συστήματα εξαγωγής πληροφορίας με χρήση πιθανοτικής ψηφοφορίας. Βασίζεται, όμως, σε μια υπόθεση η οποία περιορίζει τη χρηστικότητα της μεθόδου και αφορά την υποχρεωτική ύπαρξη ενός μόνο παραδείγματος πεδίου (OPD) σε μια σελίδα. Επιπλέον, οι υπάρχουσες τεχνικές συνδυασμού ταξινομητών δεν είναι άμεσα εφαρμόσιμες στο πρόβλημα της εξαγωγής πληροφορίας, το οποίο δεν είναι από τη φύση του ένα πρόβλημα κοινής ταξινόμησης [109].

Η παρούσα διατριβή επιχειρεί να καλύψει το κενό που υπάρχει στο χώρο του συνδυασμού συστημάτων εξαγωγής πληροφορίας, προτείνοντας μια νέα μεθοδολογία η οποία επιτρέπει το συνδυασμό πληθώρας συστημάτων, ανεξάρτητα από το πώς αυτά μοντελοποιούν το πρόβλημα της εξαγωγής πληροφορίας. Η προτεινόμενη μεθοδολογία επιτρέπει το συνδυασμό διαφορετικών συστημάτων εξαγωγής πληροφορίας, και συγκεκριμένα προσαρμοστικών συστημάτων, σε μετα-επίπεδο με χρήση τεχνικών ψηφοφορίας και συσσωρευμένης γενίκευσης, ενώ γίνεται σύγκριση και ανάλυση των αποτελεσμάτων όλων των τεχνικών, συμπεριλαμβανομένης και της πολυστρατηγικής μάθησης που περιγράφεται στην εργασία [48]. Η αποτελεσματικότητα των τεχνικών ψηφοφορίας και συσσωρευμένης γενίκευσης αξιολογείται εμπειρικά σε πληθώρα θεματικών περιοχών για να διαπιστωθεί εάν επιτυγχάνονται καλύτερα αποτελέσματα στην εξαγωγή πληροφορίας σε μετα-επίπεδο.

Η βασική υπόθεση που υιοθετείται σε αυτή τη διατριβή είναι ο συνδυασμός *διαφορετικών* συστημάτων εξαγωγής πληροφορίας τα οποία εκπαιδεύονται με εφαρμογή *διαφορετικών* αλγορίθμων μάθησης σε κοινές συλλογές αρχείων κειμένου. Στην Ενότητα 2.2, όμως, αναφέρθηκαν οι τεχνικές *ενδυνάμωσης (boosting)* και *εμφωλίασης (bagging)* κατά τις οποίες συνδυάζονται με χρήση ψηφοφορίας, διαφορετικοί ταξινομητές που προκύπτουν όμως με εφαρμογή *του ίδιου* αλγορίθμου μηχανικής μάθησης σε διαφορετικές εκδόσεις του ίδιου συνόλου των δεδομένων εκπαίδευσης (ομογενείς ταξινομητές). Αντίστοιχα, διαφορετικά συστήματα εξαγωγής πληροφορίας μπορούν να προκύψουν με την εφαρμογή του ίδιου αλγορίθμου σε διαφορετικές εκδόσεις των δεδομένων εκπαίδευσης τα οποία στη συνέχεια θα μπορούσαν να συνδυαστούν. Η διαφορά είναι βέβαια ότι σε κοινά προβλήματα ταξινόμησης υπάρχουν διανύσματα χαρακτηριστικών, ενώ στην εξαγωγή πληροφορίας υπάρχουν κείμενα επισημειωμένα με παραδείγματα πεδίων για μια θεματική περιοχή.

Αυτή η ανομοιογένεια μεταξύ των δεδομένων εκπαίδευσης και επαλήθευσης προκαλεί μια σειρά από προφανείς δυσκολίες στην άμεση εφαρμογή της εμφωλίασης και της ενδυνάμωσης στην εξαγωγή πληροφορίας, καθώς και στην θεωρητική ανάλυση των τεχνικών αυτών, η οποία ανάλυση υφίσταται για κλασσικά προβλήματα ταξινόμησης. Για παράδειγμα, η μέθοδος της ενδυνάμωσης απαιτεί τη χρήση βαρών στα δεδομένα εκπαίδευσης, κάτι που είναι άμεσα εφαρμόσιμο για ένα διάνυσμα χαρακτηριστικών σε κοινά προβλήματα ταξινόμησης. Στην εξαγωγή πληροφορίας όμως, δεν είναι προφανής η αξιοποίηση βαρών σε ένα ολόκληρο κείμενο που είναι επισημειωμένο με σχετικά παραδείγματα διαφορετικών πεδίων. Επίσης, οι διαθέσιμοι αλγόριθμοι εξαγωγής πληροφορίας δεν χειρίζονται εξ ορισμού βάρη στα επισημειωμένα παραδείγματα. Η ενδυνάμωση μπορεί να εφαρμοστεί με έμμεσο τρόπο στην εξαγωγή πληροφορίας, όταν αυτή μετατραπεί σε μια σειρά προβλημάτων κοινής ταξινόμησης [49]. Όπως, όμως έχει ήδη αναφερθεί, η μετατροπή αυτή δεν είναι ένας φυσικός τρόπος μοντελοποίησης του προβλήματος της εξαγωγής πληροφορίας γενικότερα [109].

Όσον αφορά την εμφωλίαση, σε αυτή τη διατριβή πραγματοποιήθηκαν πειράματα με ψηφοφορίες πολλαπλών συστημάτων εξαγωγής πληροφορίας που προέκυψαν από την εφαρμογή του ίδιου αλγορίθμου σε διαφορετικές εκδόσεις των δεδομένων εκπαίδευσης. Τα αποτελέσματα όμως δεν ήταν ενθαρρυντικά καθώς οι διαφορετικές εκδόσεις των συστημάτων δεν διαφέρουν σημαντικά στις προβλέψεις τους, ώστε να ευνοηθεί ο συνδυασμός και να δώσει ενθαρρυντικά αποτελέσματα. Από τη άλλη πλευρά, αυτή η διατριβή βασίζεται στις διαφορετικές *κλίσεις (biases)* που έχουν διαφορετικοί αλγόριθμοι

εξαγωγής πληροφορίας, αισιοδοξώντας στη δημιουργία συστημάτων τα οποία θα έχουν ένα ικανοποιητικό βαθμό διαφορετικότητας στις προβλέψεις που θα μπορεί να αξιοποιηθεί κατά το συνδυασμό τους σε μετα-επίπεδο, είτε με χρήση ψηφοφορίας είτε με χρήση συσσωρευμένης γενίκευσης.

Τέλος, η Ενότητα 2.3.3 αναφέρει ότι πέρα από τεχνικές ψηφοφορίας και συσσωρευμένης γενίκευσης, συνδυασμός ταξινομητών μπορεί να γίνει και με χρήση *διαδοχικής γενίκευσης*, όπου σε κάθε βήμα της ακολουθιακής διαδικασίας εκπαίδευσης, προστίθενται στο σώμα των διανυσμάτων εκπαίδευσης χαρακτηριστικά που αντιστοιχούν στις προβλέψεις του ταξινομητή του προηγούμενου επιπέδου. Τα μειονεκτήματα όμως της διαδοχικής γενίκευσης, όπως περιγράφονται στην Ενότητα 2.3.3 γίνονται μεγαλύτερα στην περίπτωση της εξαγωγής πληροφορίας, εξαιτίας της ύπαρξης επισημειωμένων κειμένων αντί διανυσμάτων χαρακτηριστικών. Συγκεκριμένα, οι προβλέψεις ενός συστήματος εξαγωγής πληροφορίας δεν αντιστοιχούν σε επιπλέον διανύσματα χαρακτηριστικών αλλά σε τμήματα κειμένου που αναγνωρίζονται ως σχετικά παραδείγματα κάποιων πεδίων. Η πληροφορία αυτή θα πρέπει να κωδικοποιηθεί κατάλληλα μέσα στα κείμενα ώστε να μπορεί να γίνει εκμεταλλεύσιμη από το επόμενο σύστημα που θα εκπαιδευτεί. Επίσης η κωδικοποίηση αυτή θα επιβαρύνει τη διαστατικότητα των κειμένων εκπαίδευσης οδηγώντας αναπόφευκτα σε μεγάλη αύξηση του υπολογιστικού κόστους. Τέτοια επιβάρυνση δεν υφίσταται στην περίπτωση της ψηφοφορίας και συσσωρευμένης γενίκευσης, όπου ένα σύνολο συστημάτων εξαγωγής εκπαιδεύονται *παράλληλα* στο ίδιο σώμα επισημειωμένων κειμένων, ενώ μόνο οι προβλέψεις τους συνδυάζονται σε μετα-επίπεδο.

ΚΕΦΑΛΑΙΟ 3

ΑΞΙΟΛΟΓΗΣΗ ΣΥΣΤΗΜΑΤΩΝ ΒΑΣΙΚΟΥ ΕΠΙΠΕΔΟΥ

Στο κεφάλαιο αυτό περιγράφονται οι θεματικές περιοχές ενδιαφέροντος, οι αλγόριθμοι που χρησιμοποιήθηκαν σε βασικό και σε μετα-επίπεδο ενώ αξιολογούνται τα συστήματα εξαγωγής πληροφορίας του βασικού επιπέδου. Η σύγκριση των συστημάτων είναι ισότιμη, καθώς πραγματοποιείται με βάση μια κοινή μεθοδολογία αξιολόγησης. Από την άλλη πλευρά, οι ερευνητές στο χώρο της εξαγωγής πληροφορίας συγκρίνουν μέχρι τώρα τα αποτελέσματα των συστημάτων τους με τα αντίστοιχα άλλων ερευνητών, δίχως όμως να εξασφαλίζονται ισότιμες συνθήκες σύγκρισης. Για παράδειγμα, δεν εξασφαλιζόταν η ύπαρξη κοινών κειμένων εκπαίδευσης και επαλήθευσης για τα συγκρινόμενα συστήματα.

Οι Ενότητες 3.1 έως 3.3 περιγράφουν τις θεματικές περιοχές και τους αλγορίθμους που αξιολογήθηκαν, ενώ η Ενότητα 3.4 περιγράφει τη μεθοδολογία αξιολόγησης. Τέλος, η Ενότητα 3.5 αναλύει τα αποτελέσματα των συστημάτων εξαγωγής πληροφορίας του βασικού επιπέδου, ενώ παράλληλα διερευνάται εάν υπάρχουν περιθώρια βελτίωσης σε μετα-επίπεδο. Τα συμπεράσματα του κεφαλαίου αυτού συνοψίζονται στην Ενότητα 3.6, ενώ τα αποτελέσματα αξιολόγησης των συστημάτων θα χρησιμοποιηθούν ως κατώφλι για τη σύγκριση όλων των τεχνικών συνδυασμού που προτείνονται σε αυτή τη διατριβή.

3.1 Θεματικές περιοχές ενδιαφέροντος

Στα πλαίσια της διατριβής αυτής πραγματοποιήθηκαν εκτενή πειράματα σε πέντε συλλογές κειμένων, επισημειωμένων με παραδείγματα σχετικών πεδίων από πέντε διαφορετικές θεματικές περιοχές αντίστοιχα. Οι τρεις συλλογές ανήκουν στο χώρο του παγκοσμίου ιστού, ενώ οι υπόλοιπες δύο ανήκουν στον ευρύτερο χώρο του διαδικτύου και συγκεκριμένα στο χώρο των *ομάδων συζητήσεων (newsgroups)*. Όλες οι συλλογές είναι διαθέσιμες στην ερευνητική κοινότητα.

Η πρώτη συλλογή αποτελείται από 101 αρχεία κειμένου τα οποία περιγράφουν πανεπιστημιακά μαθήματα της επιστήμης υπολογιστών. Τρία σχετικά πεδία επισημειώθηκαν στα αρχεία κειμένου για την θεματική αυτή περιοχή: ο κωδικός του μαθήματος, *crsnumber* (για παράδειγμα “CS302”), το όνομα του μαθήματος, *crstitle* (για παράδειγμα “Operating Systems”) και τέλος, ο διδάσκων του μαθήματος, *crsinst*, που συμπεριλαμβάνει και τυχόν βοηθούς διδασκαλίας στο μάθημα.

Η δεύτερη συλλογή αποτελείται από 96 αρχεία κειμένου τα οποία περιγράφουν ερευνητικά προγράμματα. Δύο σχετικά πεδία έχουν επισημειωθεί στα αρχεία κειμένου γι' αυτή την περιοχή: ο τίτλος ερευνητικού προγράμματος, *projtitle* (για παράδειγμα "WebKB") και το μέλος ερευνητικού προγράμματος, *projmember*. Οι δύο συλλογές σελίδων που αντιστοιχούν στις θεματικές περιοχές των μαθημάτων της επιστήμης υπολογιστών και ερευνητικών προγραμμάτων συγκεντρώθηκαν και επισημειώθηκαν στα πλαίσια του ερευνητικού προγράμματος "WebKB" [30].

Η τρίτη συλλογή αποτελείται από 50 αρχεία κειμένου τα οποία περιέχουν περιγραφές προϊόντων φορητών ηλεκτρονικών υπολογιστών. Τα αρχεία αυτά συγκεντρώθηκαν από περίπου 25 δικτυακούς τόπους. Ένα σύνολο από 19 σχετικά πεδία έχουν επισημειωθεί στα αρχεία κειμένου γι' αυτή την περιοχή τα οποία φαίνονται στον Πίνακα 3.1, μαζί με μια σύντομη περιγραφή για κάθε πεδίο. Η συλλογή αυτή σχηματίστηκε στα πλαίσια του ερευνητικού προγράμματος CROSSMARC[1].

Πίνακας 3.1 Σύντομη περιγραφή σχετικών πεδίων για τη θεματική περιοχή των φορητών ηλεκτρονικών υπολογιστών.

Πεδίο	**Σύντομη περιγραφή**
batterylife	Διάρκεια της μπαταρίας του φορητού, π.χ.. *3.5 h*
batterytype	Τύπος μπαταρίας, π.χ. Li-Ion
cdromspeed	Ταχύτητα συσκευής CD, π.χ. *40x*
dvdspeed	Ταχύτητα συσκευής DVD, π.χ. *12x*
HDcapacity	Χωρητικότητα σκληρού δίσκου, π.χ. *40GB*
manuf	Κατασκευαστής φορητού, π.χ. *Toshiba*
model	Όνομα μοντέλου του φορητού
modemSpeed	Ταχύτητα του μόντεμ, π.χ. *56k*
preinstOS	Προ-εγκατεστημένο λειτουργικό σύστημα, π.χ. *WinXP*
preinstSW	Προ-εγκατεστημένο λογισμικό, π.χ. *MS Office*
price	Τιμή φορητού, π.χ. 1200 €
procName	Επεξεργαστής του φορητού, π.χ. *Intel Pentium III*
procSpeed	Ταχύτητα επεξεργαστή, π.χ. *600 MHz*
ram	Χωρητικότητα μνήμης RAM, π.χ. *256 MB*
screenRes	Ανάλυση οθόνης, π.χ. 12*80x1024*
screenSize	Μέγεθος οθόνης, π.χ. 15''
screenType	Τύπος οθόνης, π.χ. *TFT*
warranty	Εγγύηση του φορητού, π.χ. *3 years*
weight	Βάρος του φορητού, e.g. *3 kg*

Η τέταρτη συλλογή αποτελείται από 300 αρχεία κειμένου, προερχόμενα από την ομάδα συζητήσεων *austin.jobs* στο πανεπιστήμιο του Austin στο Texas των Η.Π.Α. Τα αρχεία αυτά περιγράφουν αγγελίες εργασίας κι έχουν επισημειωθεί με 17 σχετικά πεδία, όπως

[1] CROSSMARC, R&D project, IST-2000-25366, http://www.iit.demokritos.gr/skel/crossmarc

ο τίτλος της διαθέσιμης εργασίας, ο μισθός, το όνομα της εταιρείας, κ.α. Η συλλογή αυτή έχει χρησιμοποιηθεί συχνά για αξιολόγηση στο χώρο της εξαγωγής πληροφορίας.

Η τελευταία συλλογή αποτελείται από 485 αρχεία προερχόμενα από ομάδα συζητήσεων του πανεπιστημίου Carnegie Mellon των Η.Π.Α., και περιγράφουν ανακοινώσεις σεμιναρίων. Συνολικά 4 πεδία έχουν επισημειωθεί στα αρχεία αυτά: η ώρα έναρξης του σεμιναρίου, *stime*, η ώρα λήξης του σεμιναρίου, *etime*, το όνομα του ομιλητή του σεμιναρίου, *speaker*, και τέλος, η τοποθεσία όπου θα γίνει το σεμινάριο, *location*. Επίσης η συλλογή αυτή έχει χρησιμοποιηθεί αρκετά συχνά από ερευνητές για αξιολόγηση στο χώρο της εξαγωγής πληροφορίας.

Πρέπει να τονιστεί ότι τα διαθέσιμα σχεδιότυπα για τη θεματική περιοχή των αγγελιών εργασίας δεν περιέχουν πληροφορία που αφορά τα όρια έναρξης και τέλους των επισημειωμένων παραδειγμάτων σχετικών πεδίων στον πίνακα λεκτικών μονάδων κάθε αρχείου κειμένου. Γι' αυτό, ολόκληρη η συλλογή επισημειώθηκε εκ νέου, έχοντας τα ήδη υπάρχοντα σχεδιότυπα ως οδηγό, ώστε τα καινούρια να περιέχουν πληροφορία για τα όρια έναρξης και λήξης, δηλαδή να είναι στη μορφή του Πίνακα 2.2(β).

Τέλος, για τις τρεις θεματικές περιοχές που ανήκουν στο χώρο του παγκοσμίου ιστού, οι ετικέτες HTML δεν αγνοούνται κατά την εκπαίδευση των συστημάτων εξαγωγής πληροφορίας του βασικού επιπέδου αλλά συμμετέχουν κανονικά στον πίνακα λεκτικών μονάδων κάθε σελίδας του ιστού. Για παράδειγμα, η ακολουθία χαρακτήρων *<td valign="top">* αντιστοιχεί στην υπο-ακολουθία *"td_start_tag", "attrib_valign", "value_top"*, στον πίνακα *λεκτικών μονάδων (tokens)* μιας σελίδας του ιστού.

3.2 Αλγόριθμοι σε βασικό επίπεδο

Η ενότητα αυτή περιγράφει συνοπτικά τα τρία συστήματα που αξιολογήθηκαν σε βασικό επίπεδο, προέρχονται από τρεις διαφορετικές περιοχές μηχανικής μάθησης και είναι αρκετά γνωστά στο χώρο της εξαγωγής πληροφορίας.

3.2.1 Σύστημα βασισμένο στον αλγόριθμο BWI

Η περιληπτική περιγραφή του αλγορίθμου *Boosted Wrapper Induction (BWI)* σε αυτή την ενότητα, ακολουθεί την ορολογία που υπάρχει στην εργασία [49]. Ο αλγόριθμος BWI εφαρμόζει τον αλγόριθμο ενδυνάμωσης *AdaBoost* [51] για το πρόβλημα της εξαγωγής πληροφορίας. Συγκεκριμένα, για κάθε σχετικό πεδίο μιας θεματικής περιοχής, γίνεται εκμάθηση ενός *wrapper* $W = (F, A, H)$, όπου $F = \{F^1, F^2 ...\}$ είναι ένα σύνολο κανόνων εξαγωγής πληροφορίας οι οποίοι καλούνται και *μπροστινοί ανιχνευτές (fore*

detectors), οι οποίοι αναγνωρίζουν το αρχικό όριο έναρξης των σχετικών παραδειγμάτων πεδίων στον πίνακα λεκτικών μονάδων του αρχείου κειμένου, και $A = \{A^1, A^2 ...\}$ το αντίστοιχο σύνολο των *πίσω ανιχνευτών (aft detectors)*, οι οποίοι αναγνωρίζουν το όριο τέλους των σχετικών παραδειγμάτων πεδίων. Τέλος, το H είναι ένα ιστόγραμμα μήκους των σχετικών παραδειγμάτων πεδίων, όπου $H(k)$ είναι η πιθανότητα ότι ένα σχετικό παράδειγμα πεδίου έχει μήκος k λεκτικών μονάδων. Κάθε ανιχνευτής είναι ένα ζεύγος $< pref, suff >$ ακολουθιών λεκτικών μονάδων (π.χ. <*MB, μνήμη ram*>), επαυξημένων ενδεχομένως και με τυπογραφική πληροφορία (δηλαδή αν μια λεκτική μονάδα αποτελείται από πεζούς ή κεφαλαίους ή αριθμητικούς χαρακτήρες, κλπ.). Η τυπογραφική αυτή πληροφορία είναι το απλούστερο και το μοναδικό είδος γλωσσικής πληροφορίας που εκμεταλλεύεται ο αλγόριθμος BWI κατά τη διαδικασία μάθησης. Ένα όριο λεκτικής μονάδας αναγνωρίζεται από έναν ανιχνευτή εάν η ακολουθία $pref$ των μονάδων ταιριάζει αντίστοιχο αριθμό μονάδων πριν από το όριο και η ακολουθία $suff$ ταιριάζει αντίστοιχο αριθμό μονάδων μετά από το όριο.

Για την εκμάθηση ενός συνόλου F μπροστινών ανιχνευτών για κάθε σχετικό πεδίο, η ακόλουθη μεθοδολογία εφαρμόζεται η οποία βασίζεται στον αλγόριθμο *AdaBoost*: σε όλα τα δυνατά όρια λεκτικών μονάδων που μπορούν να απαριθμηθούν σε όλα τα αρχεία κειμένου του σώματος εκπαίδευσης δίνεται ένα αρχικό βάρος το οποίο είναι ίδιο για όλα τα όρια. Σε ένα αρχείο κειμένου που αποτελείται από P λεκτικές μονάδες, υπάρχουν P όρια έναρξης και P όρια τέλους λεκτικών μονάδων. Τα όρια έναρξης και τέλους των παραδειγμάτων κάθε σχετικού πεδίου ανήκουν στο αντίστοιχο σύνολο των παραδειγμάτων έναρξης και τέλους για το πεδίο. Όλα τα υπόλοιπα όρια θεωρούνται ως αρνητικά παραδείγματα μάθησης για το πεδίο. Σε κάθε γύρο της διαδικασίας ενδυνάμωσης (100 γύροι ενδυνάμωσης χρησιμοποιούνται συνήθως για κάθε σχετικό πεδίο αλλά ο αριθμός μπορεί να είναι μεγαλύτερος), ένας *αλγόριθμος αδύναμης μάθησης (weak learner)*, αναζητά τον καλύτερο μπροστινό ανιχνευτή. Ο καλύτερος ανιχνευτής, βάσει μιας μετρικής αξιολόγησης, επιστρέφεται τελικά από τον αλγόριθμο αδύναμης μάθησης, προστίθεται στο σύνολο F, που αρχικά είναι κενό, αφού προηγουμένως καταχωρηθεί σε αυτόν ένας βαθμός εμπιστοσύνης σύμφωνα με την 3.1.

$$C(F) = 0{,}5 \ln\left(\frac{WT^+ + \varepsilon}{WT^- + \varepsilon}\right) \tag{3.1}$$

όπου WT^+ είναι το άθροισμα των βαρών των ορίων λεκτικών μονάδων που έχουν σωστά αναγνωριστεί (ταξινομηθεί), ενώ WT^- είναι το αντίστοιχο άθροισμα για τα όρια

λεκτικών μονάδων που έχουν λανθασμένα αναγνωριστεί ως σχετικά. Τέλος, ε είναι μια μικρή παράμετρος εξομάλυνσης. Στο τέλος κάθε γύρου ενδυνάμωσης, τα βάρη των ορίων λεκτικών μονάδων σε όλα τα αρχεία εκπαίδευσης, αναπροσαρμόζονται ώστε να μειωθεί το βάρος των ορίων που έχουν σωστά αναγνωριστεί ως σχετικά και να αυξηθεί το βάρος των ορίων που έχουν λανθασμένα αναγνωριστεί. Ανάλογη διαδικασία χρησιμοποιείται για την εκμάθηση του αντίστοιχου συνόλου των πίσω ανιχνευτών.

Για να πραγματοποιηθεί εξαγωγή πληροφορίας από ένα αρχείο κειμένου κατά τη διαδικασία επαλήθευσης, απαριθμούνται όλα τα δυνατά ζεύγη (s,e) ορίων λεκτικών μονάδων (συνήθως μέχρι ένα μέγιστο μήκος, π.χ. 10). Δοθέντος ενός *wrapper* W που έχει μαθευτεί για ένα σχετικό πεδίο f καταχωρείται στα όρια s,e ενός τμήματος κειμένου $t(s,e)$ ένα αρχικό κι ένα τελικό σκορ αντίστοιχα, με βάση τα σύνολα F και A των μπροστινών και πίσω ανιχνευτών και χρησιμοποιώντας τις εξισώσεις 3.2 και 3.3.

$$F(s) = \sum_k C(F^k)F^k \tag{3.2}$$

$$A(e) = \sum_k C(A^k)A^k \tag{3.3}$$

όπου $C(F^k)$ και $C(A^k)$ είναι οι βαθμοί εμπιστοσύνης από τους ανιχνευτές F^k και A^k αντίστοιχα. Εάν ένας ανιχνευτής αποτύχει να αναγνωρίσει ένα όριο, τότε ο αντίστοιχος βαθμός εμπιστοσύνης είναι μηδέν. Στη συνέχεια υπολογίζεται η τιμή

$$W(s,e) = F(s)A(e)H(e-s) \tag{3.4}$$

όπου οι ποσότητες $F(s)$ και $A(e)$ ορίστηκαν ήδη στις εξισώσεις 3.2 και 3.3, ενώ το $H(e-s)$ είναι η πιθανότητα ότι το τμήμα $t(s,e)$ είναι σχετικό, με μήκος λεκτικών μονάδων ίσο με e-s και η οποία πιθανότητα έχει υπολογιστεί κατά τη διαδικασία εκπαίδευσης. Η τιμή με βάση την εξίσωση 3.4 αντιστοιχεί και στον συνολικό βαθμό εμπιστοσύνης που εκχωρείται στο παράδειγμα $< t(s,e), f >$. Εάν ο βαθμός αυτός είναι θετικός, δηλαδή εάν $W(s,e) > 0$ τότε το αντίστοιχο παράδειγμα καταχωρείται στο τελικό σχεδιότυπο για το αρχείο από το οποίο εξάγουμε πληροφορία. Είναι προφανές ότι όσο περισσότεροι ανιχνευτές αναγνωρίσουν ένα τμήμα $t(s,e)$ ως σχετικό για ένα πεδίο f, τόσο μεγαλύτερος θα είναι ο βαθμός εμπιστοσύνης που εκχωρείται στο $< t(s,e), f >$.

Αναλυτική περιγραφή του αλγορίθμου BWI υπάρχει στην εργασία [49], ενώ μια περιεκτική ανάλυση της απόδοσης του BWI για πληθώρα προβλημάτων εξαγωγής πληροφορίας υπάρχει στην εργασία [62].

3.2.2 Σύστημα βασισμένο στον αλγόριθμο $(LP)^2$

Ο αλγόριθμος $(LP)^2$ αποτελεί συντομογραφία του *Learning Pattern by Language Processing* και παρουσιάστηκε από τον Ciravegna [24]. Ο $(LP)^2$ υλοποιεί έναν αλγόριθμο *ακολουθιακής κάλυψης (sequential covering)* για την εκμάθηση κανόνων εξαγωγής πληροφορίας. Αυτό σημαίνει ότι η μάθηση πραγματοποιείται σε ένα μη προκαθορισμένο αριθμό βημάτων ξεκινώντας από ένα σύνολο θετικών παραδειγμάτων εκπαίδευσης, δηλαδή από ένα σύνολο επισημειωμένων παραδειγμάτων πεδίων στο σώμα των κειμένων εκπαίδευσης. Σε κάθε βήμα μαθαίνεται ένα σύνολο κανόνων το οποίο αξιολογείται στο σύνολο των θετικών παραδειγμάτων πεδίων. Όσα παραδείγματα αναγνωρίζονται σε κάθε βήμα από τους κανόνες που μαθαίνονται κατά το βήμα αυτό, απομακρύνονται από το σύνολο των παραδειγμάτων εκπαίδευσης. Η διαδικασία τερματίζεται όταν καλυφθούν από τους κανόνες όλα τα παραδείγματα εκπαίδευσης.

Οι κανόνες που μαθαίνει ο $(LP)^2$ είναι *κανόνες εισαγωγής ετικετών (tagging rules)* οι οποίοι εισάγουν XML ετικέτες έναρξης και τέλους ενός παραδείγματος σχετικού πεδίου μέσα στο κείμενο. Για παράδειγμα, οι ετικέτες <*ram*> και </*ram*> εισάγονται μέσα σε ένα αρχείο κειμένου από τους κανόνες που έχουν μαθευτεί για το πεδίο *ram*, για να δηλώσουν την έναρξη και το τέλος αντίστοιχα των σχετικών παραδειγμάτων για το πεδίο αυτό που αναγνωρίζονται μέσα στο κείμενο.

Για κάθε παράδειγμα σχετικού πεδίου, ο $(LP)^2$ δημιουργεί έναν αρχικό κανόνα, στη συνέχεια γενικεύει τον κανόνα αυτό κρατώντας τις k καλύτερες γενικεύσεις σε κάθε βήμα (η τιμή του k είναι προκαθορισμένη σε κάθε βήμα) με βάση μια ειδική μετρική αξιολόγησης. Κάθε αρχικός κανόνας είναι μια σύζευξη συνθηκών σε μια ακολουθία από $2*w$ λεκτικές μονάδες, όπου οι πρώτες w μονάδες είναι στα αριστερά των αρχικών (ή τελικών) ορίων των επισημειωμένων παραδειγμάτων πεδίων και οι υπόλοιπες w μονάδες βρίσκονται στα δεξιά των αρχικών (η τελικών) ορίων των ίδιων παραδειγμάτων πεδίων. Στα πειράματα που πραγματοποιήθηκαν, η τιμή για το w τέθηκε ίση με 5. Η διαδικασία γενίκευσης κατά τη μάθηση πραγματοποιείται χαλαρώνοντας τις συνθήκες στην αρχική σύζευξη των λεκτικών μονάδων, επιτρέποντας τη χρήση γλωσσικής πληροφορίας, όπως τυπογραφική πληροφορία (για παράδειγμα εάν λεκτική μονάδα αποτελείται από πεζούς ή κεφαλαίους ή αριθμητικούς χαρακτήρες, κλπ.), πληροφορία που αφορά το μέρος του λόγου (για παράδειγμα εάν μια λεκτική μονάδα είναι ουσιαστικό ή ρήμα, κλπ.) ή πληροφορία που αφορά το λήμμα ή το στέμμα μιας μονάδας. Η διαδικασία μάθησης τερματίζεται όταν καλυφθούν από τους κανόνες όλα τα θετικά παραδείγματα πεδίων στο σώμα εκπαίδευσης.

Εκτός από τη μάθηση με χρήση ακολουθιακής κάλυψης, ο $(LP)^2$ μαθαίνει ένα επιπρόσθετο σύνολο *κανόνων περιεχομένου (contextual rules)* οι οποίοι βελτιώνουν την απόδοση στην εξαγωγή πληροφορίας των κανόνων που έχουν μαθευτεί με βάση την ακολουθιακή κάλυψη. Οι νέοι κανόνες εκμεταλλεύονται τις αλληλεπιδράσεις μεταξύ των ετικετών που εισάγονται στο κείμενο. Για παράδειγμα, η ετικέτα *</processorname>* που δηλώνει το τέλος ενός παραδείγματος του πεδίου *processorname* (όνομα επεξεργαστή) μπορεί να χρησιμοποιηθεί ως "άγκυρα" για την εισαγωγή της ετικέτας *<processorspeed>* που δηλώνει την έναρξη ενός παραδείγματος για το πεδίο *processorspeed* (ταχύτητα επεξεργαστή), αφού η ταχύτητα ενός επεξεργαστή εμφανίζεται στο κείμενο αμέσως μετά το όνομα του επεξεργαστή. Τέλος, ακολουθεί η μάθηση ενός τρίτου συνόλου κανόνων που διορθώνουν τα λάθη των προηγούμενων.

Κατά τη διαδικασία επαλήθευσης, οι κανόνες που έχουν μαθευτεί για κάθε σχετικό πεδίο f εφαρμόζονται σε ένα αρχείο κειμένου. Το περιεχόμενο $t(s,e)$ ανάμεσα σε μια ετικέτα έναρξης και μια ετικέτα τέλους (για παράδειγμα *<ram>* και *</ram>*) εξάγεται τελικά και το παράδειγμα $< t(s,e), f >$ εισάγεται στο σχεδιότυπο για τη σελίδα. Στο παράδειγμα αυτό εκχωρείται ένας βαθμός εμπιστοσύνης $LS = wrong / matched$, όπου $wrong$ είναι ο αριθμός των λαθών, κατά τη διαδικασία εκπαίδευσης, των κανόνων που αναγνώρισαν το παράδειγμα $< t(s,e), f >$, και $matched$ είναι ο συνολικός αριθμός των παραδειγμάτων που έχουν αναγνωρίσει οι κανόνες αυτοί κατά τη διαδικασία εκπαίδευσης. Είναι προφανές ότι όσο μικρότερο είναι το LS, τόσο μεγαλύτερος είναι ο βαθμός εμπιστοσύνης που εκχωρείται στο $< t(s,e), f >$. Μια πιο αναλυτική περιγραφή του $(LP)^2$ μπορεί να βρεθεί στις εργασίες [24, 26].

3.2.3 Σύστημα βασισμένο στα Κρυφά Μαρκοβιανά μοντέλα

Η μοντελοποίηση με βάση τα *Κρυφά Μαρκοβιανά μοντέλα (HMMs)* είναι μια ισχυρή στατιστική τεχνική μάθησης, κατάλληλη για τη μοντελοποίηση *ακολουθιακών δεδομένων (sequential data)*, έχοντας αντικρύσει μεγάλο εύρος εφαρμογής στις περιοχές του γραπτού και προφορικού λόγου, καθώς και σε εφαρμογές μηχανικής μετάφρασης.

Ένα Κρυφό Μαρκοβιανό μοντέλο (HMM) είναι ένα *αυτόματο πεπερασμένης κατάστασης (finite state automaton - FSA)*, με στοχαστικές (πιθανοτικές) μεταβάσεις καταστάσεων και στοχαστικές παραγωγές συμβόλων [91]. Δοθέντος ενός Κρυφού Μαρκοβιανού μοντέλου που αποτελείται από ένα σύνολο O καταστάσεων κι ένα λεξικό V διακεκριμένων συμβόλων, το μοντέλο πιθανοτικά παράγει μια ακολουθία $I = i^1 ... i^Z$ συμβόλων, όπου κάθε σύμβολο $i^k \in V$, ξεκινώντας από μια αρχική κατάσταση $o^1 \in O$,

παράγοντας πιθανοτικά ένα σύμβολο i^1, μεταβαίνοντας επίσης πιθανοτικά στην επόμενη κατάσταση $o^2 \in O$, παράγοντας ξανά ένα δεύτερο σύμβολο i^2 και συνεχίζοντας μέχρι την παραγωγή και του τελευταίου συμβόλου i^z.

Σε ένα πρόβλημα αναγνώρισης προτύπων, η διαδικασία αυτή μπορεί να φορμαλιστεί ως μια διαδικασία εύρεσης της βέλτιστης ακολουθίας καταστάσεων $\hat{O} = o^1 ... o^Z$ η οποία αντιστοιχεί στην ακολουθία συμβόλων $I = i^1 ... i^Z$, δηλαδή

$$\hat{O} = \arg\max_O P(O \mid I) = \arg\max_O \frac{P(O) \cdot P(I \mid O)}{P(I)} \tag{3.5}$$

όπου $P(O \mid I)$ είναι η πιθανότητα της ακολουθίας καταστάσεων O, δοθείσας της ακολουθίας συμβόλων I. Εξαιτίας του γεγονότος ότι το πρόβλημα βελτιστοποίησης που περιγράφει η εξίσωση 3.5 είναι ανεξάρτητο του I, και λαμβάνοντας υπόψη τη μαρκοβιανή υπόθεση πρώτης τάξης, τότε το πρόβλημα περιορίζεται στην εύρεση του

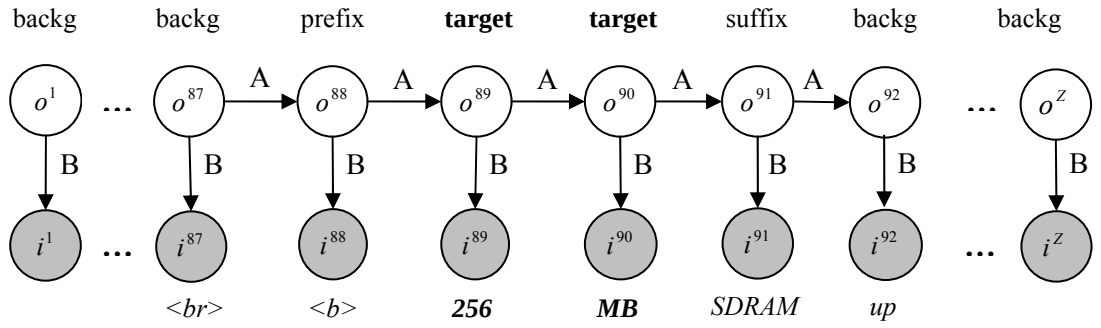
$$\arg\max_{o^1 ... o^Z} \left(\prod_{k=1}^{Z} P(o^k \mid o^{k-1}) \cdot P(i^k \mid o^k) \right) \tag{3.6}$$

κάτι που μπορεί να υπολογιστεί με τη βοήθεια του αλγορίθμου *Viterbi* [115].

Για το πρόβλημα της εξαγωγής πληροφορίας, η χρήση των HMMs γίνεται ως εξής [50, 100]: κάθε σελίδα κειμένου θεωρείται ότι παράγεται στοχαστικά από ένα HMM, ξεκινώντας από μια αρχική κατάσταση του μοντέλου όπου παράγεται στοχαστικά η πρώτη λεκτική μονάδα του κειμένου, μεταβαίνοντας στη συνέχεια επίσης στοχαστικά στην επόμενη κατάσταση όπου παράγεται η δεύτερη λεκτική μονάδα και συνεχίζοντας μέχρι να παραχθεί και η τελευταία λεκτική μονάδα του κειμένου. Επίσης, ένα HMM παράγεται για κάθε σχετικό πεδίο μιας θεματικής περιοχής, το λεξικό V του οποίου αποτελείται από τις διακεκριμένες λεκτικές μονάδες που έχουν απαριθμηθεί σε όλα τα αρχεία εκπαίδευσης, ενώ το σύνολο καταστάσεων O ορίζεται ως εξής: ένα σύνολο καταστάσεων που καλούνται “target” καταστάσεις, παράγουν τις λεκτικές μονάδες που έχουν επισημειωθεί από τον ειδικό της θεματικής περιοχής ως σχετικά τμήματα κειμένου για το πεδίο που εξετάζεται. Ένα σύνολο από “prefix” καταστάσεις παράγουν τις μονάδες που υπάρχουν αμέσως πριν από εκείνες που αντιστοιχούν σε σχετικά τμήματα για το πεδίο. Ένα άλλο σύνολο “suffix” καταστάσεων παράγουν τις μονάδες που υπάρχουν αμέσως μετά από εκείνες των σχετικών τμημάτων κειμένου. Όλες οι υπόλοιπες λεκτικές μονάδες που υπάρχουν σε ένα αρχείο κειμένου δεν είναι σχετικές

για το πεδίο και παράγονται από μια ειδική κατάσταση η οποία καλείται “backg”. Το Σχήμα 3.1 δείχνει το γραφικό μοντέλο ενός HMM, σε μορφή άκυκλου κατευθυνόμενου γράφου, το οποίο μοντελοποιεί παραδείγματα του πεδίου *ram*, σύμφωνα με το δείγμα σελίδας του Πίνακα 2.2(α).

Για το HMM του Σχήματος 3.1 έχει γίνει η υπόθεση ότι $O = \{$*backg*, *prefix*, *suffix*, *target*$\}$, δηλαδή ότι το μοντέλο αποτελείται από 4 διακεκριμένες καταστάσεις, όπου κάθε $o^k \in O$ είναι η ετικέτα της κατάστασης που αντιστοιχεί στη λεκτική μονάδα $i^k \in V$ που παράγεται. Πρέπει να παρατηρηθεί ότι το τμήμα κειμένου “256 MB” που αποτελείται από δύο λεκτικές μονάδες, παράγεται από την ίδια κατάσταση “target” κάτι που σημαίνει αυτόματα την ύπαρξης μια *αυτό-μετάβασης (self-transition)* για την κατάσταση αυτή.



Σχήμα 3.1 Ένα HMM, σε μορφή άκυκλου κατευθυνόμενου γράφου για παραδείγματα του πεδίου *ram* στη θεματική περιοχή των φορητών υπολογιστών, όπου $o^k \in O = \{$*backg*, *prefix*, *target*, *suffix*$\}$, *k* μια χρονικά μεταβαλλόμενη μεταβλητή, $A = P(o^k \mid o^{k-1})$ ο πίνακας των πιθανοτήτων μετάβασης καταστάσεων και, $B = P(i^k \mid o^k)$ ο πίνακας με τις πιθανότητες παραγωγής συμβόλων.

Ένα καινοτομικό στοιχείο της χρήσης HMMs για εξαγωγή πληροφορίας, σε αντίθεση με τη χρήση τους για προβλήματα αναγνώρισης φωνής, είναι η ένα-προς-ένα αντιστοιχία μεταξύ των καταστάσεων του μοντέλου κι ενός συνόλου ετικετών για τις καταστάσεις αυτές. Θεωρούμε δηλαδή ότι κάθε διακεκριμένη κατάσταση του μοντέλου έχει μια συγκεκριμένη σημασία (για παράδειγμα “target, “suffix”, “prefix”, backg”). Επομένως, η ακολουθία καταστάσεων δεν είναι πλέον “κρυφή” και έτσι δεν απαιτείται η χρήση του αλγορίθμου Baum-Welch για τον υπολογισμό των παραμέτρων του μοντέλου. Οι παράμετροι αυτοί είναι ο $N \times N$ πίνακας A πιθανοτήτων μετάβασης καταστάσεων (N ο αριθμός των καταστάσεων του μοντέλου), ο $N \times M$ πίνακας B πιθανοτήτων παραγωγής συμβόλων (M ο αριθμός των διακεκριμένων συμβόλων) κι ένας $N \times 1$ πίνακας αρχικοποίησης που περιέχει την πιθανότητα για κάθε κατάσταση να παράγει την πρώτη λεκτική μονάδα κάθε σελίδας.

Οι παράμετροι ενός HMM για κάθε σχετικό πεδίο υπολογίζονται με βάση τα επισημειωμένα κείμενα εκπαίδευσης και με χρήση απλών πράξεων μέτρησης των μεταβάσεων καταστάσεων και της παραγωγής λεκτικών μονάδων από αυτές τις καταστάσεις. Κάποια τεχνική *εξομάλυνσης (smoothing)* εφαρμόζεται στη συνέχεια για την αποφυγή πιθανοτήτων με μηδενική τιμή. Μια αρκετά γνωστή τεχνική εξομάλυνσης που εφαρμόζεται συχνά ονομάζεται *Witten-Bell smoothing* [118].

Δοθέντος ενός αρχείου κειμένου κατά την επαλήθευση κι ενός HMM για ένα σχετικό πεδίο f, ο *Viterbi* βρίσκει τη βέλτιστη ακολουθία καταστάσεων $\hat{O} = o^1 ... o^Z$ που αντιστοιχεί στην ακολουθία λεκτικών μονάδων $I = i^1 ... i^Z$ του κειμένου. Στη συνέχεια, οι λεκτικές μονάδες που αντιστοιχούν στις καταστάσεις εκείνες με ετικέτα “target” εξάγονται και τα αντίστοιχα παραδείγματα $< t(s,e), f >$ εισάγονται στο σχεδιότυπο για το αρχείο κειμένου. Εάν υποθέσουμε ότι η ακολουθία καταστάσεων που φαίνεται στο πάνω μέρος του Σχήματος 3.1 έχει παραχθεί από τον *Viterbi*, δοθέντος του εκπαιδευμένου μοντέλου για το πεδίο *ram*, τότε το τμήμα κειμένου “256 MB” εξάγεται και το αντίστοιχο παράδειγμα $<"256MB", ram>$ εισάγεται στο σχεδιότυπο.

Όσον αφορά το βαθμό εμπιστοσύνης που εκχωρεί κάθε HMM ενός πεδίου στα τμήματα κειμένου που αναγνωρίζει ως σχετικά μέσω του αλγορίθμου *Viterbi*, προτείνεται σε αυτή τη διατριβή η χρήση της παρακάτω φόρμουλας 3.7:

$$\log \prod_{k=a}^{b} P(o^k \mid o^{k-1}) \cdot P(i^k \mid o^k) \qquad (3.7)$$

όπου τα a και b είναι τα όρια των λεκτικών μονάδων που παράγονται από την πρώτη “prefix” κατάσταση και την τελευταία “suffix” κατάσταση αντίστοιχα, σύμφωνα με τη βέλτιστη ακολουθία καταστάσεων που επιστρέφεται από τον *Viterbi*. Για το παράδειγμα του Σχήματος 3.1, $a = 88$ και $b = 91$. Όσο μεγαλύτερες είναι οι τιμές των πιθανοτήτων μετάβασης καταστάσεων $P(o^k \mid o^{k-1})$ και παραγωγής συμβόλων $P(i^k \mid o^k)$, όπως υπολογίστηκαν κατά την εκπαίδευση, τόσο μεγαλύτερος είναι ο βαθμός εμπιστοσύνης που καταχωρείται στο παράδειγμα του πεδίου.

Το κίνητρο για τη χρήση της φόρμουλας 3.7 περιγράφεται ως εξής: δοθέντος του HMM, το γινόμενο των πιθανοτήτων μεταβάσεων καταστάσεων και των πιθανοτήτων παραγωγής λεκτικών μονάδων, παράγεται από τον *Viterbi* αλγόριθμο, όπως φαίνεται και από την εξίσωση 3.6. Χρησιμοποιώντας την 3.7, το γινόμενο της 3.6 περικόπτεται γύρω από τις σχετικές με την εξαγωγή καταστάσεις (με ετικέτες “prefix”, “suffix”,

"target"). Η χρήση του λογαρίθμου στην 3.7 επιβάλλεται για το χειρισμό πολύ μικρών πιθανοτήτων, όπως τυπικά είναι οι πιθανότητες μετάβασης καταστάσεων και παραγωγής συμβόλων. Χρησιμοποιώντας την 3.7 διασφαλίζεται το γεγονός ότι τμήματα κειμένου που είναι σχετικά με ένα πεδίο και εντοπίζονται σε διαφορετικά μέρη ενός αρχείου κειμένου, θα αντιστοιχούν σε διαφορετικές τιμές εμπιστοσύνης και όχι σε μια ενιαία τιμή όπως επιστρέφει ο *Viterbi* για ολόκληρο το αρχείο κειμένου.

Πρέπει βέβαια να τονιστεί ότι λόγω του γινομένου στη φόρμουλα 3.7, τμήματα κειμένου με μεγαλύτερο μήκος (δηλαδή μεγαλύτερο αριθμό λεκτικών μονάδων) που είναι σχετικά με ένα πεδίο αντιστοιχούν σε χαμηλότερες τιμές εμπιστοσύνης. Κάτι τέτοιο συμβαδίζει με τη φυσική αντίληψη ότι τμήματα κειμένου με μικρότερο μήκος, είναι πιθανότερο να είναι σχετικά για ένα πεδίο και γι' αυτό θα πρέπει να καταχωρούνται σε μεγαλύτεροι βαθμοί εμπιστοσύνης. Το [91] είναι ένα άριστο εισαγωγικό άρθρο στη θεωρία των HMMs, ενώ λεπτομέρειες για την εφαρμογή τους σε προβλήματα εξαγωγής πληροφορίας υπάρχουν στις εργασίες [50, 100, 102].

3.3 Αλγόριθμοι σε μετα-επίπεδο

Για εξαγωγή πληροφορίας σε μετα-επίπεδο αξιολογήθηκαν οι παρακάτω αλγόριθμοι μηχανικής μάθησης, οι οποίοι είναι σχεδιασμένοι για προβλήματα ταξινόμησης και έχουν υλοποιηθεί στα πλαίσια της πλατφόρμας WEKA [119].

- *J48*, που αποτελεί υλοποίηση του αρκετά γνωστού αλγορίθμου C4.5 [90] εκμάθησης δέντρων απόφασης.
- *NaiveBayes,* που αποτελεί υλοποίηση του γνωστού αλγορίθμου Naïve Bayes [60].
- *IB1,* που αποτελεί υλοποίηση του αλγορίθμου του 1-κοντινότερου γείτονα.
- *Multi-response Linear Regression-MLR,* που είναι υλοποίηση του αλγορίθμου πολύ-ανταποκριτικής γραμμικής παλινδρόμησης, ο οποίος είναι προσαρμογή της *γραμμικής παλινδρόμησης ελάχιστων τετραγώνων (least-squares linear regression, Breiman 1996a)* για τον μετασχηματισμό ενός προβλήματος ταξινόμησης σε ένα πρόβλημα παλινδρόμησης.
- *SMO*, που αποτελεί μια γρήγορη υλοποίηση του γνωστού αλγορίθμου *Support Vector Machines-SVM* [88].
- *LogitBoost,* που αποτελεί υλοποίηση του αντίστοιχου αλγορίθμου [52], όπου επίσης ένα πρόβλημα ταξινόμησης μετασχηματίζεται σε ένα πρόβλημα (λογιστικής) παλινδρόμησης.

Πρέπει να τονιστεί ότι πληθώρα άλλων αλγορίθμων θα μπορούσαν να αξιολογηθούν σε μετα-επίπεδο. Η επιλογή των συγκεκριμένων αλγορίθμων έγινε αφενός διότι οι αλγόριθμοι αυτοί έχουν χρησιμοποιηθεί στη βιβλιογραφία για την αξιολόγηση της συσσωρευμένης γενίκευσης σε κοινά προβλήματα ταξινόμησης [41, 98, 107] και αφετέρου διότι οι αλγόριθμοι αυτοί καλύπτουν διαφορετικές περιοχές μάθησης.

Το WEKA είναι μια πλατφόρμα εξόρυξης γνώσης από δεδομένα που περιέχει μια ευρεία συλλογή αλγορίθμων μηχανικής μάθησης, υλοποιημένων στη γλώσσα προγραμματισμού JAVA, ενώ παρέχει και γραφικό περιβάλλον τόσο για τη χρήση τους, όσο και για την προ-επεξεργασία των δεδομένων εκπαίδευσης και αξιολόγησης, καθώς και για την ανάλυση των αποτελεσμάτων μέσω *οπτικοποίησης (visualization)*. Το WEKA διατίθεται δωρεάν για ερευνητικούς σκοπούς *(http://www.cs.waikato.ac.nz/~ml/weka)*.

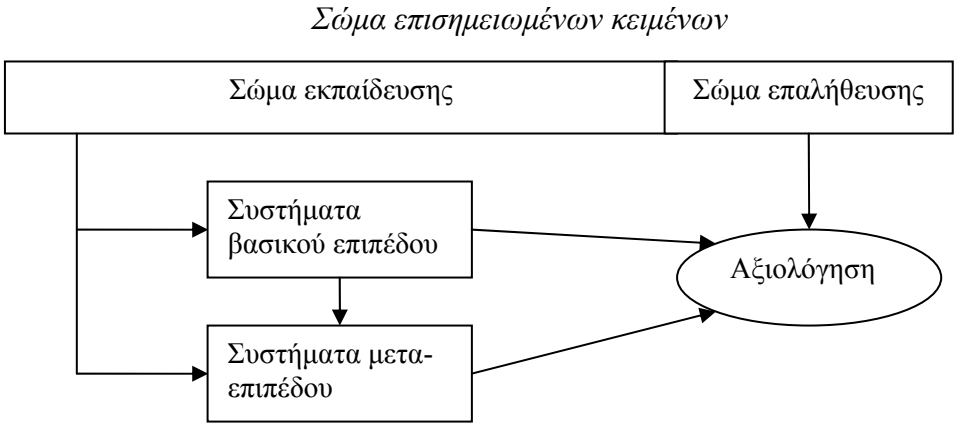
3.4 Μεθοδολογία αξιολόγησης

Για την αξιολόγηση, τόσο σε βασικό όσο και σε μετα-επίπεδο, χρησιμοποιήθηκε διασταυρωμένη επικύρωση στο σύνολο των επισημειωμένων κειμένων για κάθε θεματική περιοχή. Για την περιοχή των φορητών ηλεκτρονικών υπολογιστών, το σώμα των 50 αρχείων κειμένου χωρίστηκε με τυχαία επιλογή σε 5 υποσύνολα των 10 αρχείων. Σε καθένα από τα 5 βήματα της διασταυρωμένης επικύρωσης, τα 40 αρχεία χρησιμοποιούνται για την εκπαίδευση των συστημάτων του βασικού επιπέδου και των ταξινομητών σε μετα-επίπεδο, ενώ τα υπόλοιπα 10 αρχεία χρησιμοποιούνται για την αξιολόγηση. Τέλος, τα αποτελέσματα συγκεντρώνονται και για τα 5 βήματα.

Για την περιοχή των αγγελιών εργασιών, επίσης μια διεργασία διασταυρωμένης επικύρωσης 5 βημάτων πραγματοποιήθηκε για την αξιολόγηση. Συγκεκριμένα, το σώμα των 300 αρχείων χωρίστηκε με τυχαία επιλογή σε 5 υποσύνολα των 60 αρχείων. Σε καθένα από τα 5 βήματα της διασταυρωμένης επικύρωσης, τα 240 αρχεία χρησιμοποιούνται για την εκπαίδευση των συστημάτων του βασικού επιπέδου και των ταξινομητών του μετα-επιπέδου, ενώ τα υπόλοιπα 60 αρχεία χρησιμοποιούνται για την αξιολόγηση. Τέλος, τα αποτελέσματα συγκεντρώνονται και για τα 5 βήματα.

Για τις περιοχές των μαθημάτων της επιστήμης υπολογιστών, των ερευνητικών προγραμμάτων και των ανακοινώσεων σεμιναρίων, μια διαφορετική μεθοδολογία αξιολόγησης ακολουθήθηκε, για την επίτευξη αντικειμενικής αξιολόγησης με τα αποτελέσματα στην εργασία [48]. Για κάθε μια από τις τρεις αυτές περιοχές, το σώμα των αρχείων κειμένου χωρίστηκε σε δύο υποσύνολα περίπου του ίδιου μεγέθους. Το πρώτο μέρος χρησιμοποιήθηκε για την εκπαίδευση των συστημάτων του βασικού

επιπέδου και των ταξινομητών του μετα-επιπέδου, ενώ το δεύτερο μέρος χρησιμοποιήθηκε για την αξιολόγηση. Μια εσωτερική διαδικασία διασταυρωμένης επικύρωσης τριών βημάτων ακολουθήθηκε στο υποσύνολο εκπαίδευσης (πρώτο μέρος) για τη δημιουργία των διανυσμάτων χαρακτηριστικών και την εκπαίδευση των ταξινομητών σε μετα-επίπεδο. Η όλη διαδικασία επαναλήφθηκε 5 φορές, με τα τελικά αποτελέσματα να συγκεντρώνονται στο τέλος. Το Σχήμα 3.2 δείχνει σχηματικά τη διαδικασία αξιολόγησης για ένα βήμα της διασταυρωμένης επικύρωσης.



Σχήμα 3.2 Αξιολόγηση των συστημάτων εξαγωγής πληροφορίας σε βασικό και σε μετα-επίπεδο, σε ένα βήμα της διαδικασίας διασταυρωμένης επικύρωσης. Τα τελικά αποτελέσματα αξιολόγησης συγκεντρώνονται από όλα τα βήματα.

Κατά την αξιολόγηση, ο περιορισμός του *ενός παραδείγματος ανά κείμενο* (OPD) εφαρμόστηκε για τα πεδία *crsNumber* και *crsTitle* στην περιοχή των μαθημάτων της επιστήμης υπολογιστών, για το πεδίο *projTitle* στα ερευνητικά προγράμματα και για όλα τα 4 πεδία στις ανακοινώσεις σεμιναρίων. Αυτό σημαίνει ότι μόνο ένα στιγμιότυπο από τα πεδία αυτά μπορεί να υπάρχει σε ένα αρχείο κειμένου. Στην περίπτωση περισσοτέρων από ένα παραδείγματα για ένα OPD πεδίο σε ένα αρχείο, το παράδειγμα εκείνο με τη μεγαλύτερη πιθανότητα κρατείται και εισάγεται στο τελικό σχεδιότυπο για τη σελίδα κειμένου, ενώ τα υπόλοιπα απορρίπτονται. Η αξιολόγηση πραγματοποιείται με τη σύγκριση των χειρονακτικά συμπληρωμένων σχεδιοτύπων και των αντίστοιχων που έχουν συμπληρωθεί από τα συστήματα εξαγωγής πληροφορίας.

Τρεις μετρικές χρησιμοποιήθηκαν για αξιολόγηση των συστημάτων εξαγωγής πληροφορίας, τόσο σε βασικό όσο και σε μετα-επίπεδο: η *ακρίβεια (precision-P)* που αντιστοιχεί στο πηλίκο των σωστών παραδειγμάτων πεδίων που έχουν αναγνωριστεί από το σύστημα προς τα συνολικά παραδείγματα που έχουν αναγνωριστεί, η *ανάκληση (recall-R)* που αντιστοιχεί στο πηλίκο των σωστών παραδειγμάτων πεδίων που έχουν

αναγνωριστεί, προς το σύνολο των σωστών παραδειγμάτων πεδίων (που υπάρχουν στα χειρονακτικά συμπληρωμένα σχεδιότυπα), και τέλος η μετρική $F1$ που είναι ο αρμονικός μέσος της ακρίβειας και της ανάκλησης, σύμφωνα και με την 3.8.

$$F1 = \frac{2 * R * P}{R + P} \tag{3.8}$$

Ο ορισμός της ακρίβειας και της ανάκλησης όπως ορίζεται σε αυτή την ενότητα, είναι ισοδύναμος με τη *μικρο-αθροιστική (micro-average)* ακρίβεια και τη μικρο-αθροιστική ανάκληση [95] τα οποία ορίζονται επίσημα ως εξής:

$$P = \frac{TP}{TP + FP} = \frac{\sum_{I=1}^{Q} TP^I}{\sum_{I=1}^{Q} (TP^I + FP^I)} \tag{3.9}$$

$$R = \frac{TP}{TP + FN} = \frac{\sum_{I=1}^{Q} TP^I}{\sum_{I=1}^{Q} (TP^I + FN^I)} \tag{3.10}$$

όπου TP^i είναι ο αριθμός των παραδειγμάτων του πεδίου $f^i \in \{f^1, ... f^Q\}$ που έχουν σωστά αναγνωριστεί *(σωστά θετικά/true positive)*, FP^i είναι ο αριθμός των παραδειγμάτων του πεδίου f^i που έχουν λανθασμένα ταξινομηθεί *(λανθασμένα αρνητικά/false negative)*, και FN^i είναι ο αριθμός των παραδειγμάτων του πεδίου f^i που δεν έχουν αναγνωριστεί *(λανθασμένα αρνητικά/false negative)*. Η επιλογή των μικρο-αθροιστικών μετρικών τιμών επιτρέπει την αντικειμενικότερη αξιολόγηση διαφορετικών συστημάτων εξαγωγής, αφού οι μετρικές αυτές υπολογίζονται με βάση όλα τα παραδείγματα των σχετικών πεδίων μιας θεματικής περιοχής.

Η *στατιστική σημαντικότητα (statistical significance)* των συγκρίσεων που πραγματοποιούνται σε αυτή τη διατριβή αξιολογήθηκε σύμφωνα το αρκετά γνωστό τεστ *paired t-test* [38] με ποσοστό σημαντικότητας το 95%.

3.5 Αξιολόγηση σε βασικό επίπεδο

Το πρόβλημα της μη ισότιμης σύγκρισης συστημάτων εξαγωγής πληροφορίας έχει επισημανθεί πρόσφατα στη βιβλιογραφία [75]. Κάθε ερευνητής ακολουθεί μια διαφορετική διαδικασία διασταυρωμένης επικύρωσης κατά την οποία γίνεται διαχωρισμός του συνόλου των επισημειωμένων κειμένων για μια θεματική περιοχή σε κείμενα εκπαίδευσης και κείμενα επαλήθευσης. Η διατριβή αυτή προσφέρει μια ισότιμη σύγκριση των τριών συστημάτων εξαγωγής του βασικού επιπέδου σε πέντε θεματικές

περιοχές, χρησιμοποιώντας κοινή μεθοδολογία αξιολόγησης, δηλαδή κοινά σύνολα κειμένων εκπαίδευσης και κειμένων επαλήθευσης για κάθε σύστημα.

Η Ενότητα 3.5.1 παρουσιάζει και αναλύει τα αποτελέσματα των συστημάτων εξαγωγής πληροφορίας του βασικού επιπέδου. Η Ενότητα 3.5.2 επιχειρεί μια ποσοτική μέτρηση της διαφορετικότητας της εξόδου των συστημάτων αυτών ώστε να διαπιστωθεί εάν υπάρχουν περιθώρια βελτίωσης της απόδοσης της εξαγωγής σε μετα-επίπεδο. Επίσης απουσιάζει από τη βιβλιογραφία παρόμοια μέτρηση διαφορετικότητας στην έξοδο των συγκεκριμένων συστημάτων και σε ανάλογο εύρος περιοχών.

3.5.1 Παρουσίαση και ανάλυση αποτελεσμάτων

Οι Πίνακες 3.2 έως 3.6 δείχνουν τα αποτελέσματα που επιτυγχάνονται στις πέντε θεματικές περιοχές αντίστοιχα από τα τρία συστήματα εξαγωγής πληροφορίας που είναι διαθέσιμα σε βασικό επίπεδο. Οι Πίνακες 3.7 έως 3.9 συγκρίνουν τα τρία συστήματα του βασικού επιπέδου με βάση τον αριθμό των στατιστικά σημαντικότερων *νικών* έναντι *ηττών* στις πέντε περιοχές, χρησιμοποιώντας τις τρεις μετρικές αξιολόγησης αντίστοιχα. Τέλος, ο Πίνακας 3.10 πραγματοποιεί την ίδια σύγκριση αλλά σε κάθε σχετικό πεδίο χωριστά και για τις πέντε περιοχές και χρησιμοποιώντας μόνο τη μετρική $F1$. Τα Παραρτήματα Α.1 έως και Α.3 δείχνουν αναλυτικές τιμές για όλα τα πεδία.

Πίνακας 3.2 Αποτελέσματα σε βασικό επίπεδο για τα μαθήματα της επιστήμης υπολογιστών.

Σύστημα του βασικού επιπέδου	Ακρίβεια	Ανάκληση	F1
BWI	74.55	39.10	51.30
HMM	60.50	58.29	59.39
$(LP)^2$	71.39	60.90	65.73

Πίνακας 3.3 Αποτελέσματα σε βασικό επίπεδο για τα ερευνητικά προγράμματα.

Σύστημα του βασικού επιπέδου	Ακρίβεια	Ανάκληση	F1
BWI	60.05	61.47	60.75
HMM	56.24	68.18	61.64
$(LP)^2$	63.31	54.92	58.82

Πίνακας 3.4 Αποτελέσματα σε βασικό επίπεδο για τους φορητούς ηλεκτρονικούς υπολογιστές.

Σύστημα του βασικού επιπέδου	Ακρίβεια	Ανάκληση	F1
BWI	74.99	53.23	62.26
HMM	62.29	65.42	63.81
$(LP)^2$	63.24	59.41	61.26

Πίνακας 3.5 Αποτελέσματα σε βασικό επίπεδο για τις αγγελίες εργασίας.

Σύστημα του βασικού επιπέδου	Ακρίβεια	Ανάκληση	F1
BWI	89.42	72.39	80.01
HMM	72.42	79.31	75.71
$(LP)^2$	87.70	79.18	83.22

Πίνακας 3.6 Αποτελέσματα σε βασικό επίπεδο για τις ανακοινώσεις σεμιναρίων.

Σύστημα του βασικού επιπέδου	Ακρίβεια	Ανάκληση	F1
BWI	93.26	74.92	83.09
HMM	78.34	80.09	79.20
$(LP)^2$	91.39	81.63	86.23

Πίνακας 3.7 Στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, χρησιμοποιώντας τη μετρική *ακρίβεια* στις πέντε θεματικές περιοχές ενδιαφέροντος.

	BWI	HMM	$(LP)^2$
BWI		5\0	4\1
HMM	0\5		0\4
$(LP)^2$	1\4	4\0	

Πίνακας 3.8 Στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, χρησιμοποιώντας τη μετρική *ανάκληση* στις πέντε θεματικές περιοχές ενδιαφέροντος.

	BWI	HMM	$(LP)^2$
BWI		0\5	1\4
HMM	5\0		2\1
$(LP)^2$	4\1	1\2	

Πίνακας 3.9 Στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, χρησιμοποιώντας τη μετρική $F1$ στις πέντε θεματικές περιοχές ενδιαφέροντος ενδιαφέροντος.

	BWI	HMM	$(LP)^2$
BWI		2\2	0\3
HMM	2\2		1\3
$(LP)^2$	3\0	3\1	

Πίνακας 3.10 Στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, χρησιμοποιώντας τη μετρική $F1$, στα 45 πεδία συνολικά και των πέντε θεματικών περιοχών.

	BWI	HMM	$(LP)^2$
BWI		12\13	2\20
HMM	13\12		8\21
$(LP)^2$	20\2	21\8	

Από τα αποτελέσματα που φαίνονται στους παραπάνω πίνακες βγαίνει το συμπέρασμα ότι το σύστημα εξαγωγής πληροφορίας που βασίζεται στον $(LP)^2$ είναι το καλύτερο στις περισσότερες θεματικές περιοχές, όσο και στα περισσότερα πεδία συνολικά.

Από την άλλη πλευρά, το σύστημα των HMMs είναι ιδιαίτερα ανταγωνιστικό με τα υπόλοιπα δύο συστήματα στις περιοχές των ερευνητικών προγραμμάτων και των φορητών υπολογιστών. Το συμπέρασμα αυτό είναι ιδιαίτερα ενδιαφέρον, αναλογιζόμενοι το γεγονός ότι τα HMMs, όπως είναι γνωστό και από την Ενότητα 3.2, εκμεταλλεύονται μόνο πληροφορία που αφορά τις *λεκτικές μονάδες κειμένου (tokens)*, ενώ τα υπόλοιπα δύο συστήματα εκμεταλλεύονται γλωσσική πληροφορία. Όσον αφορά τους φορητούς υπολογιστές, η περιοχή αυτή είναι περισσότερο δομημένη από τις άλλες δύο περιοχές. Η υπάρχουσα δομή συνίσταται περισσότερο στην παρουσία HTML

ετικετών (για παράδειγμα **,**,**,** κ.α.) μέσα στο κείμενο και κατά συνέπεια γύρω από τα σχετικά τμήματα κειμένου που πρέπει να αναγνωριστούν κατά την εξαγωγή. Επομένως οι κανόνες εξαγωγής που θα μαθευτούν θα πρέπει να βασίζονται και στις HTML ετικέτες για τις οποίες η γλωσσική πληροφορία που εκμεταλλεύονται τα συστήματα που βασίζονται στους BWI και $(LP)^2$ δεν είναι ιδιαίτερα χρήσιμη.

Μια επιπρόσθετη εξήγηση για την ανταγωνιστική συμπεριφορά των HMMs στα ερευνητικά προγράμματα και στους φορητούς υπολογιστές, είναι η συχνή ύπαρξη, περισσότερο από ότι ισχύει στην περιοχή των μαθημάτων της επιστήμης υπολογιστών, σχετικών τμημάτων κειμένου τόσο στα κείμενα εκπαίδευσης όσο και στα κείμενα επαλήθευσης. Για παράδειγμα, το τμήμα κειμένου “256 MB” που αντιστοιχεί σε ένα παράδειγμα του πεδίου “ram” υπάρχει επισημειωμένο αρκετά συχνά στα κείμενα εκπαίδευσης και επαλήθευσης. Επίσης, τα ονόματα κάποιων μελών ερευνητικών προγραμμάτων εμφανίζονται αρκετά συχνά τόσο στα κείμενα εκπαίδευσης όσο και σε εκείνα της επαλήθευσης. Κάτι τέτοιο διευκολύνει τα *παραγωγικά (generative)* μοντέλα μάθησης, όπως είναι τα HMMs, που αποστηθίζουν ουσιαστικά παραδείγματα πεδίων κατά την εκπαίδευση και αναγνωρίζουν τα παραδείγματα αυτά κατά την επαλήθευση. Έτσι δικαιολογείται η αρκετά ανταγωνιστική ανάκληση των HMMs, σε σχέση με τα άλλα δύο συστήματα, στα ερευνητικά προγράμματα και τους φορητούς υπολογιστές.

Αντίθετα, στην περιοχή των μαθημάτων της επιστήμης υπολογιστών, η απόδοση του συστήματος των HMMs δεν είναι ιδιαίτερα ανταγωνιστική σε σχέση με το σύστημα του $(LP)^2$. Αυτό οφείλεται στο ότι λιγότερο συχνά, σε σχέση με τις άλλες δύο περιοχές, υπάρχουν κοινά παραδείγματα πεδίων στα κείμενα εκπαίδευσης και επαλήθευσης.

Παρατηρώντας επίσης τους Πίνακες 3.7 έως 3.9, συμπεραίνουμε ότι το σύστημα του BWI επιτυγχάνει πιο ακριβή αποτελέσματα στην εξαγωγή πληροφορίας από τα άλλα δύο συστήματα στις περισσότερες θεματικές περιοχές. Από την άλλη πλευρά, το σύστημα των HMMs επιτυγχάνει αποτελέσματα με μεγαλύτερη ανάκληση στις περισσότερες θεματικές περιοχές. Όμως σύμφωνα με τη μετρική $F1$, που είναι ο αρμονικός μέσος της ακρίβειας και της ανάκλησης, το σύστημα του $(LP)^2$ είναι το καλύτερο στις περισσότερες θεματικές περιοχές ενδιαφέροντος.

Τέλος, από τους Πίνακες 3.2 έως 3.10 φαίνεται ότι κανένα σύστημα δεν είναι καθολικά καλύτερο σε όλες τις θεματικές περιοχές ή σε όλα τα πεδία μιας περιοχής. Η παρατήρηση αυτή καθιστά ακόμα πιο επιτακτική την ανάγκη αναζήτησης κάποιας μεθόδου συνδυασμού των συστημάτων του βασικού επιπέδου ώστε να επιτευχθούν καλύτερα αποτελέσματα στην εξαγωγή πληροφορίας σε μετα-επίπεδο.

3.5.2 Μέτρηση διαφορετικότητας σε βασικό επίπεδο

Παρατηρώντας ξανά του Πίνακες 3.2 έως 3.6, φαίνεται ότι η πιο απλοϊκή λύση που είναι διαθέσιμη είναι η επιλογή του καλύτερου συστήματος για κάθε θεματική περιοχή, δηλαδή η επιλογή του συστήματος $(LP)^2$ για τα μαθήματα της επιστήμης υπολογιστών, των αγγελιών εργασίας και των ανακοινώσεων σεμιναρίων και των HMMs για τις περιοχές των ερευνητικών προγραμμάτων και των φορητών υπολογιστών. Μια πιο αποτελεσματική λύση, με βάση και τον Πίνακα 3.8, θα ήταν για παράδειγμα η επιλογή διαφορετικών συστημάτων για διαφορετικά πεδία. Όμως σε ένα πρόβλημα εξαγωγής πληροφορίας δε μπορεί να είναι γνωστό εκ των προτέρων ποιο σύστημα είναι καλύτερο και για ποια πεδία. Πρέπει να τονιστεί ότι χρησιμοποιείται περισσότερο η μετρική $F1$ για την αξιολόγηση διαφορετικών συστημάτων, αφού ισορροπεί εξίσου μεταξύ ακρίβειας και ανάκλησης και έχει χρησιμοποιηθεί αρκετά συχνά σε προηγούμενες έρευνες τόσο στην εξαγωγή όσο και στην *ανάκτηση πληροφορίας (information retrieval)*.

Από την άλλη πλευρά, μια πιο επιθυμητή επιλογή θα ήταν να προσπαθήσουμε να εκμεταλλευτούμε τη διαφορετικότητα στις προβλέψεις των τριών συστημάτων του βασικού επιπέδου, ελπίζοντας να βελτιώσουμε σε μετα-επίπεδο την απόδοσή τους στην εξαγωγή. Είναι, όμως εφικτή μια τέτοια βελτίωση; Υπάρχει δηλαδή επαρκής διαφορετικότητα στις προβλέψεις των συστημάτων του βασικού επιπέδου ώστε να μπορέσουμε να εκμεταλλευτούμε τη διαφορετικότητα αυτή σε μετα-επίπεδο; Στη βιβλιογραφία δεν υπάρχει κάποιος γενικά αποδεκτός ορισμός για τη διαφορετικότητα και το πώς αυτή μετριέται. Οι Ali και Pazzani [5] όρισαν την ομοιότητα μεταξύ δύο ταξινομητών ως την υπο-συνθήκη πιθανότητα ότι και οι δύο ταξινομητές κάνουν λάθος σε μια πρόβλεψη, δοθέντος του γεγονότος ότι ο ένας από τους δύο κάνει λάθος.

Οι Πίνακες 3.11(α) έως 3.11(ε) είναι παραδείγματα *πινάκων ενδεχομένων (contingency tables)*, όπως ορίστηκαν από τον Freitag [48] για τη μέτρηση της ομοιότητας σε ζευγάρια συστημάτων εξαγωγής πληροφορίας και είναι εμπνευσμένα από τη μετρική των Ali και Pazzani [5]. Οι πίνακες συμπληρώθηκαν με βάση την έξοδο των τριών συστημάτων σε κάθε περιοχή κατά την αξιολόγηση.

Κάθε κελί σε έναν πίνακα ενδεχομένων μετράει την υπο-συνθήκη πιθανότητα ότι το σύστημα που αντιστοιχεί η γραμμή προβλέπει σωστά, δοθέντος του γεγονότος ότι και το σύστημα στήλης προβλέπει επίσης σωστά. Σωστή πρόβλεψη για ένα σύστημα σημαίνει ότι είτε προβλέπεται το σωστό πεδίο για ένα σχετικό τμήμα κειμένου, είτε ότι δεν προβλέπεται κάποιο πεδίο (αγνοούμενη πρόβλεψη) για ένα τμήμα κειμένου που έχει αναγνωριστεί λανθασμένα ως σχετικό από άλλο σύστημα ή άλλα συστήματα.

Πίνακας 3.11 Κάθε κελί μετράει την πιθανότητα ότι το σύστημα εξαγωγής πληροφορίας που αντιστοιχεί σε γραμμή κάνει μια σωστή πρόβλεψη, δοθέντος του ότι το σύστημα που αντιστοιχεί σε στήλη προβλέπει επίσης σωστά.

	BWI	HMM	$(LP)^2$
BWI	1	0.46	0.44
HMM	0.70	1	0.52
$(LP)^2$	0.89	0.69	1

(α) Μαθήματα επιστήμης υπολογιστών

	BWI	HMM	$(LP)^2$
BWI	1	0.82	0.79
HMM	0.90	1	0.79
$(LP)^2$	0.69	0.62	1

(β) Ερευνητικά προγράμματα

	BWI	HMM	$(LP)^2$
BWI	1	0.73	0.78
HMM	0.89	1	0.83
$(LP)^2$	0.87	0.76	1

(γ) Φορητοί υπολογιστές

	BWI	HMM	$(LP)^2$
BWI	1	0.83	0.86
HMM	0.91	1	0.85
$(LP)^2$	0.93	0.85	1

(δ) Αγγελίες εργασίας

	BWI	HMM	$(LP)^2$
BWI	1	0.87	0.83
HMM	0.91	1	0.84
$(LP)^2$	0.95	0.92	1

(ε) Ανακοινώσεις σεμιναρίων

Οι τιμές στους Πίνακες 3.11(α) έως 3.11(ε) δείχνουν ότι υπάρχει περιθώριο βελτίωσης στο καλύτερο, βάσει του $F1$, σύστημα εξαγωγής για κάθε θεματική περιοχή. Για παράδειγμα, στα μαθήματα της επιστήμης υπολογιστών παρατηρούμε ότι μόνο στο 69% όπου το σύστημα των HMMs κάνει μια σωστή πρόβλεψη, το σύστημα του $(LP)^2$ προβλέπει επίσης σωστά. Αυτό σημαίνει αυτόματα ότι στο υπόλοιπο 31% όπου τα HMMs προβλέπουν σωστά, ο $(LP)^2$ κάνει λανθασμένες προβλέψεις. Επομένως, η απόδοση στην εξαγωγή πληροφορίας του $(LP)^2$, που είναι το καλύτερο σύστημα για τη συγκεκριμένη περιοχή, μπορεί να βελτιωθεί περαιτέρω σε μετα-επίπεδο. Αντίστοιχα συμπεράσματα προκύπτουν και για τις υπόλοιπες τέσσερις θεματικές περιοχές.

Παρατηρούμε βέβαια ότι στις θεματικές περιοχές των αγγελιών εργασίας και των ανακοινώσεων σεμιναρίων, το περιθώριο βελτίωσης σε μετα-επίπεδο είναι μικρότερο από ότι στις άλλες τρεις περιοχές. Στις ανακοινώσεις σεμιναρίων, στο 92% των περιπτώσεων όπου το σύστημα των HMMs προβλέπει κάνει μια σωστή πρόβλεψη, το σύστημα του $(LP)^2$ προβλέπει επίσης σωστά. Δηλαδή μόλις στο υπόλοιπο 8% όπου το σύστημα των HMMs προβλέπει σωστά, το αντίστοιχο του $(LP)^2$, που είναι το καλύτερο για τη συγκεκριμένη θεματική περιοχή, κάνει λανθασμένες προβλέψεις. Θα ήταν ενδιαφέρον, λοιπόν, να διερευνηθεί η αποτελεσματικότητα τεχνικών συνδυασμού σε περιπτώσεις που είναι περιορισμένα τα περιθώρια βελτίωσης σε μετα-επίπεδο.

3.6 Συμπεράσματα

Η αξιολόγηση των τριών συστημάτων σε βασικό επίπεδο έδειξε ότι δεν υπάρχει σύστημα το οποίο να είναι καθολικά καλύτερο σε όλες τις θεματικές περιοχές. Ακόμα κι αν κάποιο σύστημα είναι καλύτερο για μια περιοχή, για κάποια επιμέρους σχετικά πεδία, άλλα συστήματα επιτυγχάνουν καλύτερα αποτελέσματα εξαγωγής. Η αξιολόγηση έδειξε πάντως ότι το σύστημα εξαγωγής του $(LP)^2$ είναι το καλύτερο σε βασικό επίπεδο. Πρέπει να σημειωθεί πάντως ότι στην εξαγωγή πληροφορίας δεν υπάρχει κάποια πλατφόρμα η οποία να περιέχει πληθώρα υλοποιήσεων αλγορίθμων, όπως υπάρχει η γνωστή πλατφόρμα WEKA [119] για ταξινόμηση. Η συλλογή αλγορίθμων για εξαγωγή πληροφορίας είναι αρκετά δυσκολότερη από τη συλλογή αλγορίθμων για ταξινόμηση.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η δυνατότητα συνδυασμού των τριών διαφορετικών συστημάτων εξαγωγής πληροφορίας που αξιολογήθηκαν σε βασικό επίπεδο. Βασικό συστατικό επιτυχίας μιας τεχνικής συνδυασμού είναι η εκμετάλλευση της διαφορετικότητας στις προβλέψεις των συστημάτων του βασικού επιπέδου, για την επίτευξη καλύτερων αποτελεσμάτων στην εξαγωγή σε μετα-επίπεδο. Η διαφορετικότητα αυτή μετρήθηκε ποσοτικά κι έδειξε ότι υπάρχουν περιθώρια βελτίωσης σε μετα-επίπεδο του καλύτερου συστήματος εξαγωγής του βασικού επιπέδου για κάθε θεματική περιοχή. Σε κάποιες περιοχές (αγγελίες εργασίας, ανακοινώσεις σεμιναρίων) βέβαια, τα περιθώρια βελτίωσης μετρήθηκαν μικρότερα σε σχέση με τις υπόλοιπες περιοχές.

Η τελευταία παρατήρηση οδηγεί στο ιδιαίτερα ενδιαφέρον κίνητρο για τη μελέτη της αποδοτικότητας τεχνικών συνδυασμού συστημάτων εξαγωγής πληροφορίας, σε περιπτώσεις όπου υπάρχει περιορισμένο περιθώριο βελτίωσης σε μετα-επίπεδο. Η διαφορετικότητα στην έξοδο των συστημάτων του βασικού επιπέδου είναι *θεμιτή* και απαραίτητη για την επιτυχία μιας τεχνικής συνδυασμού, αλλά δε μπορεί σε καμία περίπτωση να γίνει *απαιτητή*. Επιθυμούμε τα συστήματα που συνδυάζουμε να είναι αρκετά διαφορετικά στις προβλέψεις τους ώστε να βελτιώσουμε όσο το δυνατόν περισσότερο την απόδοσή τους σε μετα-επίπεδο. Σε κάποια προβλήματα, όμως, το μέγεθος της διαφορετικότητας μπορεί να μην είναι αρκετά ικανοποιητικό, όπως στις θεματικές περιοχές των ανακοινώσεων σεμιναρίων και των αγγελιών εργασίας.

Μπορούν σε τέτοιες περιπτώσεις οι τεχνικές συνδυασμού να αποδειχτούν μη ζημιογόνες, για την εξαγωγή πληροφορίας; Εναλλακτικά, είναι ωφέλιμες ή ζημιογόνες οι τεχνικές συνδυασμού σε περιπτώσεις μεγάλης ομοιότητας στην έξοδο των συστημάτων που συνδυάζονται; Το ερώτημα αυτό απαντάται σε αυτή τη διατριβή.

ΚΕΦΑΛΑΙΟ 4

ΣΥΝΔΥΑΣΜΟΣ ΣΥΣΤΗΜΑΤΩΝ ΕΞΑΓΩΓΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΜΕ ΧΡΗΣΗ ΨΗΦΟΦΟΡΙΑΣ

Στόχος του συνδυασμού συστημάτων εξαγωγής πληροφορίας είναι η εκμετάλλευση της διαφορετικότητας στις προβλέψεις τους, προσδοκώντας σε καλύτερα αποτελέσματα εξαγωγής σε μετα-επίπεδο. Η Ενότητα 4.1 περιγράφει ένα παράδειγμα συνδυασμού συστημάτων. Προτείνεται ταυτόχρονα η χρήση του *συσσωρευμένου σχεδιοτύπου*, που είναι σημαντική για το συνδυασμό συστημάτων εξαγωγής πληροφορίας.

Ο απλούστερος τρόπος συνδυασμού συστημάτων είναι με χρήση *ψηφοφορίας*. Η χρήση *πλειοψηφικής ψηφοφορίας* για εξαγωγή πληροφορίας περιγράφεται στην Ενότητα 4.2. Η χρήση *πιθανοτικής ψηφοφορίας* περιγράφεται στην Ενότητα 4.3, ενώ η Ενότητα 4.4 περιγράφει της διαφορές της πιθανοτικής ψηφοφορίας με την πολυστρατηγική μάθηση [48], τη μοναδική μέχρι τώρα σχετική εργασία στη διεθνή βιβλιογραφία όσον αφορά το συνδυασμό συστημάτων εξαγωγής πληροφορίας. Στην Ενότητα 4.5 αναλύονται τα δεδομένα σε μετα-επίπεδο, δηλαδή τα περιεχόμενα του συσσωρευμένου σχεδιοτύπου, για τον προσδιορισμό του βαθμού συσχέτισης της εξόδου των συστημάτων του βασικού επιπέδου. Η ανάλυση αυτή χρησιμοποιείται ως βάση για την ανάλυση των αποτελεσμάτων όλων των τεχνικών συνδυασμού που προτείνονται σε αυτή τη διατριβή. Στην Ενότητα 4.6 γίνεται συγκριτική αξιολόγηση των τεχνικών ψηφοφορίας, ενώ η Ενότητα 4.7 συνοψίζει τα συμπεράσματα του κεφαλαίου.

4.1 Παράδειγμα συνδυασμού διαφορετικών συστημάτων εξαγωγής πληροφορίας - Το συσσωρευμένο σχεδιότυπο

Έστω $L^1 ... L^N$ ένα σύνολο N αλγορίθμων μηχανικής μάθησης σχεδιασμένων για εξαγωγή πληροφορίας, στους οποίους δίνεται ένα σώμα D από κείμενα εκπαίδευσης, τα οποία έχουν επισημειωθεί με παραδείγματα σχετικών πεδίων μιας θεματικής περιοχής. Οι αλγόριθμοι $L^1 ... L^N$ εφαρμόζονται στο σώμα των επισημειωμένων κειμένων εκπαίδευσης προς την εκμάθηση ενός συνόλου κανόνων εξαγωγής, οι οποίοι κανόνες αναγνωρίζουν συγκεκριμένα τμήματα ως παραδείγματα σχετικών πεδίων. Έστω $E^1 ... E^N$ το αντίστοιχο σύνολο των εκπαιδευμένων συστημάτων εξαγωγής πληροφορίας τα οποία χρησιμοποιούν τους κανόνες που έχουν προκύψει για την αναγνώριση παραδειγμάτων σχετικών πεδίων σε καινούρια κείμενα. Κάθε εκπαιδευμένο σύστημα εξαγωγής πληροφορίας αποτελείται από ένα σύνολο *εννοιών στόχων (target concepts)*

που έχουν προκύψει για τα σχετικά πεδία της θεματικής περιοχής. Τέλος, έστω $T^1 ... T^N$ ένα σύνολο σχεδιοτύπων για ένα κείμενο d, συμπληρωμένα με παραδείγματα σχετικών πεδίων από τα συστήματα $E^1 ... E^N$ αντίστοιχα. Για παράδειγμα, έστω E^1, E^2 δύο συστήματα εκπαιδευμένα για εξαγωγή πληροφορίας από σελίδες του παγκοσμίου ιστού που περιγράφουν προϊόντα φορητών ηλεκτρονικών υπολογιστών.

Οι πίνακες 4.1(α) και 4.1(β) δείχνουν τα δύο σχεδιότυπα T^1, T^2 συμπληρωμένα από τα συστήματα E^1, E^2 χρησιμοποιώντας τη σελίδα του Πίνακα 2.2(α) που περιγράφει τα χαρακτηριστικά ενός φορητού υπολογιστή.

Πίνακας 4.1 Σχεδιότυπα T^1, T^2 συμπληρωμένα από δύο συστήματα εξαγωγής πληροφορίας E^1, E^2 για τη σελίδα του Πίνακα 2.2(α).

T^1		
$t(s,e)$	s, e	*Πεδίο* f
TransPort ZX	47, 49	model
15"	56, 58	screenSize
TFT	59, 60	screenType
Intel<b>Pentium III	63, 67	procName
600 MHz	67, 69	procSpeed
256 MB	76, 78	ram
1 GB	81, 83	ram

(a)

T^2		
$t(s,e)$	s, e	*Πεδίο* f
TransPort ZX	47, 49	manuf
TFT	59, 60	screenType
Intel <b> Pentium	63, 66	procName
600 MHz	67, 69	procSpeed
256 MB	76, 78	Ram
1 GB	81, 83	HDcapacity
40 GB	86, 88	HDcapacity

(b)

Εξετάζοντας προσεκτικά τους Πίνακες 4.1(α) και 4.1(β) παρατηρούμε κάποια διαφωνία στις προβλέψεις των δύο συστημάτων εξαγωγής πληροφορίας E^1, E^2. Για δύο τμήματα κειμένου (“Transport ZX” και “1GB”) τα πεδία που έχουν προβλέψει τα συστήματα E^1, E^2 διαφέρουν. Για το “Transport ZX”, το σύστημα E^1 προέβλεψε *model* (μοντέλο φορητού) ενώ το E^2 προέβλεψε *manuf* (κατασκευαστής φορητού). Για το “1GB”, το σύστημα E^1 προέβλεψε *ram* ενώ το σύστημα E^2 προέβλεψε *HDcapacity* (χωρητικότητα σκληρού δίσκου). Συγκρίνοντάς τα με το χειρονακτικά συμπληρωμένο σχεδιότυπο του Πίνακα 2.2(β), συμπεραίνουμε ότι το “Transport ZX” έχει προβλεφθεί σωστά από το σύστημα E^1 και λανθασμένα από το σύστημα E^2. Από την άλλη πλευρά, το “1 GB” δεν υπάρχει στο σχεδιότυπο του Πίνακα 2.2(β). Γι’ αυτό, τα πεδία που έχουν προβλεφθεί από τα δύο συστήματα εξαγωγής πληροφορίας είναι λανθασμένα.

Επιπλέον, κάποια τμήματα κειμένου έχουν προβλεφθεί μόνο από ένα από τα δύο συστήματα εξαγωγής πληροφορίας. Το τμήμα “15"” έχει προβλεφθεί μόνο από το σύστημα E^1 ως *screenSize* (μέγεθος οθόνης), ενώ το τμήμα “40 GB” έχει προβλεφθεί σωστά μόνο από το σύστημα E^2 ως *HDcapacity*. Τα πεδία που έχουν προβλεφθεί και στις δύο περιπτώσεις είναι σωστά. Υπάρχουν βέβαια και περιπτώσεις όπου οι

προβλέψεις και των δύο συστημάτων συμφωνούν (“TFT”, “600 MHz”, “256 MB”). Τέλος, υπάρχει και μια περίπτωση αλληλεπικαλυπτόμενων προβλέψεων από τα δύο συστήματα. Το τμήμα “Intell<b>Pentium III” έχει προβλεφθεί σωστά ως *procName* (όνομα επεξεργαστή) από το σύστημα E^1, ενώ το τμήμα “Intell<b>Pentium” έχει προβλεφθεί λανθασμένα ως *procName* από το σύστημα E^2.

Σε αυτή τη διατριβή προτείνουμε ότι η διαφωνία στις προβλέψεις των δύο συστημάτων εξαγωγής πληροφορίας E^1, E^2 μπορούν να παρατηρηθούν καλύτερα με χρήση του *συσσωρευμένου σχεδιοτύπου* του Πίνακα 4.2, το οποίο ουσιαστικά συνενώνει τα ξεχωριστά σχεδιότυπα T^1, T^2.

Πίνακας 4.2 Συσσωρευμένο σχεδιότυπο από T^1, T^2. Κάθε καταχώρηση στο συσσωρευμένο σχεδιότυπο αντιστοιχεί σε ένα τμήμα κειμένου που έχει αναγνωριστεί από τουλάχιστον ένα σύστημα εξαγωγής πληροφορίας σε βασικό επίπεδο.

s, e	$t(s,e)$	*Έξοδος από* E^1	*Έξοδος από* E^2	*Σωστό πεδίο*
47, 49	TransPort ZX	model	manuf	model
56, 58	15"	screenSize	-	screenSize
59, 60	TFT	screenType	screenType	screenType
63, 66	Intel<b>Pentium	-	procName	-
63, 67	Intel<b>Pentium III	procName	-	procName
67, 69	600 MHz	procSpeed	procSpeed	procSpeed
76, 78	256 MB	ram	ram	ram
81, 83	1 GB	ram	HDcapacity	-
86, 88	40 GB	-	HDcapacity	HDcapacity

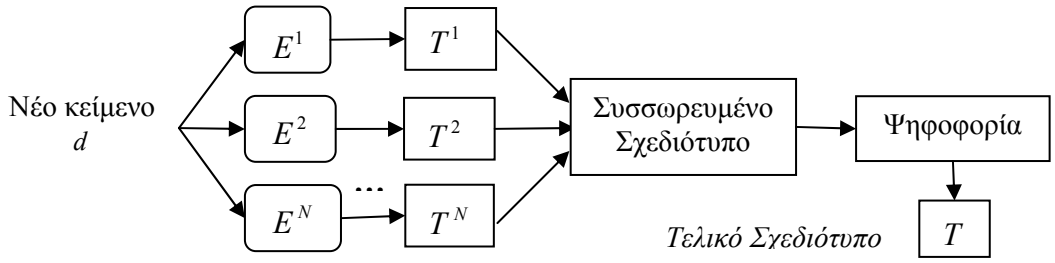
Η κατασκευή του συσσωρευμένου σχεδιοτύπου του Πίνακα 4.2 είναι μια απλοϊκή διαδικασία: όλα τα τμήματα κειμένου $t(s,e)$ που έχουν αναγνωριστεί από τα συστήματα E^1, E^2 στα σχεδιότυπα T^1, T^2 συγκεντρώνονται σε ένα σημείο, με τα διπλά αναγνωρισμένα τμήματα κειμένου να απομακρύνονται. Δύο τμήματα κειμένου διαφέρουν εάν είτε το αρχικό τους όριο είτε το τελικό τους όριο διαφέρει. Για τα εναπομείναντα *διακεκριμένα* τμήματα κειμένου, τα πεδία που έχουν προβλεφθεί από τα συστήματα E^1, E^2 συγκεντρώνονται και εισάγονται μαζί με το σωστό πεδίο (τελευταία στήλη του Πίνακα 4.2) στο συσσωρευμένο σχεδιότυπο. Εάν κάποιο σύστημα εξαγωγής πληροφορίας δεν προβλέψει κάποιο πεδίο για ένα τμήμα κειμένου, το αντίστοιχο κελί στο συσσωρευμένο σχεδιότυπο μένει κενό. Εάν ένα τμήμα κειμένου δεν υπάρχει στο χειρονακτικά συμπληρωμένο σχεδιότυπο (για παράδειγμα το “1GB”), το αντίστοιχο κελί στην τελευταία στήλη του συσσωρευμένου σχεδιοτύπου μένει επίσης κενό. Η διαδικασία αυτή μπορεί εύκολα να γενικευτεί και για N σχεδιότυπα $T^1 ... T^N$.

Το μέγεθος του συσσωρευμένου σχεδιοτύπου, δηλαδή ο αριθμός των καταχωρήσεων γραμμών, εξαρτάται από το μέγεθος των ξεχωριστών σχεδιοτύπων $T^1 ... T^N$. Στον Πίνακα 4.1, το μέγεθος και των δύο σχεδιοτύπων T^1, T^2 είναι 7, ενώ το μέγεθος του συσσωρευμένου σχεδιοτύπου στον Πίνακα 4.2 είναι 9. Αυτό οφείλεται στα 5 τμήματα κειμένου που έχουν προβλεφθεί και από τα δύο συστήματα εξαγωγής πληροφορίας E^1, E^2, όπως φαίνεται και στα αντίστοιχα σχεδιότυπα T^1, T^2, και στα 4 τμήματα κειμένου τα οποία βρίσκονται είτε στο σχεδιότυπο T^1 είτε στο T^2.

Εξετάζοντας προσεκτικά το συσσωρευμένο σχεδιότυπο του Πίνακα 4.2, αναρωτιόμαστε αν μπορούμε, σε κάποιο ανώτερο επίπεδο, να εκμεταλλευτούμε τη διαφωνία στις προβλέψεις των δύο διαφορετικών συστημάτων E^1, E^2, στοχεύοντας σε μεγαλύτερη ακρίβεια στην εξαγωγή πληροφορίας. Το επιθυμητό αποτέλεσμα είναι η αυτόματη συμπλήρωση της τελευταίας στήλης του Πίνακα 4.2 με τα σωστά πεδία. Με άλλα λόγια, θα θέλαμε να καταχωρήσουμε το σωστό πεδίο σε κάθε τμήμα κειμένου που έχει αναγνωριστεί από τουλάχιστον ένα σύστημα εξαγωγής πληροφορίας.

4.2 Συνδυασμός με χρήση πλειοψηφικής ψηφοφορίας

Μια απλή ιδέα για το συνδυασμό των προβλέψεων διαφορετικών συστημάτων εξαγωγής είναι η χρήση πλειοψηφικής ψηφοφορίας: *για κάθε γραμμή στο συσσωρευμένο σχεδιότυπο μετράμε τα πεδία που έχουν προβλεφθεί από τα διαθέσιμα συστήματα και επιλέγουμε εκείνο με το μεγαλύτερο αριθμό ψήφων*. Σε περίπτωση ισοπαλίας, τυχαία επιλογή πραγματοποιείται τυπικά ανάμεσα στα πεδία που ισοβαθμούν. Μια άλλη εναλλακτική σε περίπτωση ισοβαθμίας είναι η επιλογή του πεδίου εκείνου με τον μεγαλύτερο αριθμό επισημειωμένων παραδειγμάτων στο σώμα των κειμένων εκπαίδευσης. Το Σχήμα 4.1 δείχνει τον τρόπο συνδυασμού διαφορετικών συστημάτων εξαγωγής μέσω ψηφοφορίας.



Σχήμα 4.1 Συνδυασμός συστημάτων εξαγωγής πληροφορίας με χρήση ψηφοφορίας.

Δοθέντος ενός νέου κειμένου d, τα συστήματα $E^1 ... E^N$ εφαρμόζονται στο κείμενο και συμπληρώνουν τα αντίστοιχα σχεδιότυπα $T^1 ... T^N$. Ένα συσσωρευμένο σχεδιότυπο

δημιουργείται με βάση τα $T^1 ... T^N$, σύμφωνα με την Ενότητα 4.1, ενώ ψηφοφορία λαμβάνει χώρα τέλος, ώστε να συμπληρωθεί το τελικό σχεδιότυπο T για το κείμενο d.

Για το συσσωρευμένο σχεδιότυπο του Πίνακα 4.2, παρατηρούμε ότι αυτό περιέχει αγνοούμενες τιμές, οι οποίες αντανακλούν το φυσικό γεγονός του ότι κάποιο σύστημα δεν έχει προβλέψει κάποιο πεδίο για ένα τμήμα κειμένου που έχει αναγνωριστεί από κάποιο άλλο σύστημα. Η σπουδαιότητα των αγνοουμένων τιμών πρέπει να ληφθεί σοβαρά υπόψη. Για παράδειγμα εάν κάποιο σύστημα προβλέψει ένα λανθασμένο πεδίο f για ένα τμήμα κειμένου $t(s,e)$, ενώ τα υπόλοιπα συστήματα δεν προβλέψουν καθόλου κάποιο πεδίο για το ίδιο τμήμα, τότε παραβλέποντας τις αγνοούμενες τιμές κατά την ψηφοφορία, επηρεάζεται αρνητικά η ακρίβεια στην εξαγωγή, αφού επιστρέφεται το λανθασμένο πεδίο f. Βέβαια, εάν για ένα σύστημα γνωρίζουμε ότι οι προβλέψεις του σε μια θεματική περιοχή έχουν μεγάλη ακρίβεια, τότε ενδεχομένως μια σωστή τακτική θα ήταν να εμπιστευτούμε τις προβλέψεις αυτές, παραβλέποντας τυχόν αγνοούμενες προβλέψεις από τα υπόλοιπα συστήματα εξαγωγής πληροφορίας.

Μια εναλλακτική προσέγγιση είναι η καταγραφή κάθε αγνοούμενης πρόβλεψης για ένα τμήμα κειμένου $t(s,e)$ ως μια ειδική ονομαστική τιμή “false”, υποδεικνύοντας ότι το $t(s,e)$ είναι λανθασμένο και δεν θα έπρεπε επομένως να προβλεφθεί κανένα πεδίο για το τμήμα αυτό. Εάν η τιμή με το μεγαλύτερο αριθμό ψήφων είναι η τιμή “false”, τότε κανένα πεδίο δεν επιστρέφεται για το $t(s,e)$. Εάν όμως το πεδίο f αντιστοιχεί σε σωστή πρόβλεψη για το $t(s,e)$, τότε μεταφράζοντας τις αγνοούμενες τιμές από τα υπόλοιπα συστήματα ως “false”, βλάπτει τη συνολική απόδοση στην εξαγωγή, αφού το σωστό πεδίο f απορρίπτεται.

Γι’ αυτό, δύο διαφορετικά σχήματα πλειοψηφικής ψηφοφορίας ορίζονται και αξιολογούνται σε αυτή τη διατριβή, ανάλογα με το εάν οι αγνοούμενες τιμές παραβλέπονται ή μεταφράζονται ως τιμές “false”, υποδεικνύοντας άρνηση πρόβλεψης.

4.3 Συνδυασμός με χρήση πιθανοτικής ψηφοφορίας

Η ψηφοφορία συστημάτων εξαγωγής πληροφορίας μπορεί να πραγματοποιηθεί και με χρήση πιθανοτήτων ορθότητας στην έξοδο των συστημάτων, αντί της χρήσης απλών ονομαστικών τιμών. Η ψηφοφορία με χρήση πιθανοτήτων που παρουσιάζεται σε αυτή την ενότητα έχει αρκετά κοινά γνωρίσματα με την *πολυστρατηγική μάθηση* [48] που συνοψίζεται στην Ενότητα 2.5. Και τα δύο σχήματα μοιράζονται την ίδια μέθοδο για την αντιστοιχία βαθμού εμπιστοσύνης σε πιθανότητα ορθότητας στις προβλέψεις των

συστημάτων εξαγωγής πληροφορίας, καθώς και την ίδια Εξίσωση 2.1 η οποία υπολογίζει τη συνδυασμένη πιθανότητα για ένα παράδειγμα $< t(s,e), f >$.

Σε αυτή τη διατριβή ορίζονται δύο διαφορετικά σχήματα πιθανοτικής ψηφοφορίας. Στο πρώτο σχήμα, αγνοούμενες τιμές παραβλέπονται, όπως και στο πρώτο σχήμα της πλειοψηφικής ψηφοφορίας. Δοθέντος ενός τμήματος κειμένου $t(s,e)$, τότε καταχωρείται στο $t(s,e)$ το πεδίο f με τη μεγαλύτερη (συνδυασμένη) πιθανότητα ορθότητας από εκείνα τα συστήματα εξαγωγής που έχουν προβλέψει κάποιο πεδίο για το $t(s,e)$.

Στο δεύτερο σχήμα, όμως, ένας περιορισμός εισάγεται για το εάν το πεδίο f με τη μεγαλύτερη πιθανότητα θα πρέπει να γίνει δεκτό ή να απορριφθεί. Εάν η πιθανότητα για το f είναι μικρότερη από 0.5, τότε το f απορρίπτεται. Ειδάλλως, το f γίνεται δεκτό και το παράδειγμα $< t(s,e), f >$ εισάγεται στο τελικό σχεδιότυπο για το κείμενο. Η τιμή 0.5 είναι μια λογική επιλογή κατωφλίου για το εάν ένα πεδίο f πρέπει να γίνει δεκτό ή όχι. Το Σχήμα 4.2 περιγράφει αλγοριθμικά την προτεινόμενη διαδικασία για το συνδυασμό συστημάτων εξαγωγής πληροφορίας με χρήση πιθανοτικής ψηφοφορίας.

Διαδικασία Πιθανοτικής Ψηφοφορίας για Εξαγωγή Πληροφορίας

Είσοδος: Κείμενο d, Εκπαιδευμένα συστήματα $E^1 ... E^N$
Έξοδος: Συμπληρωμένο σχεδιότυπο T για το κείμενο d
Διαδικασία: $T^1 ... T^N$ = συμπληρωμένα σχεδιότυπα των $E^1 ... E^N$ για το d
ΣΣ = συσσωρευμένο σχεδιότυπο από τα $T^1 ... T^N$
Για κάθε γραμμή στο *ΣΣ*, δηλαδή για κάθε τμήμα κειμένου $t(s,e)$
- Βρες τα πεδία που έχουν προβλεφθεί για το $t(s,e)$
- Για κάθε διαφορετικό πεδίο f^c υπολόγισε τη συνδυασμένη πιθανότητα

$$P^c = 1 - \prod (1 - p^j)$$

όπου p^j η πιθανότητα το σύστημα E^i να προβλέπει το πεδίο f^c
- Βρες το πεδίο f^c με τη μεγαλύτερη συνδυασμένη πιθανότητα P^c

Περίπτωση 1:
- Καταχώρησε το παράδειγμα $< t(s,e), f >$ στο τελικό σχεδιότυπο T

Περίπτωση 2:
- Καταχώρησε το παράδειγμα $< t(s,e), f >$ στο τελικό σχεδιότυπο T, μόνο εάν P^c >= 0.5

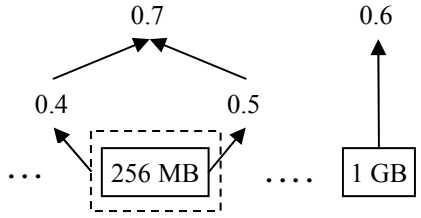
Σχήμα 4.2 Προτεινόμενη μεθοδολογία συνδυασμού συστημάτων εξαγωγής πληροφορίας μέσω ψηφοφορίας με χρήση πιθανοτήτων ορθότητας στην έξοδο των συστημάτων.

Πρέπει τα τονιστεί ότι δε μπορεί να οριστεί σχήμα πιθανοτικής ψηφοφορίας που να ερμηνεύει την απουσία πρόβλεψης ενός συστήματος ως “false” υποδεικνύοντας άρνηση πρόβλεψης, όπως στην αντίστοιχη περίπτωση της πλειοψηφικής ψηφοφορίας. Το πρόβλημα που προκύπτει στην περίπτωση αυτή είναι ότι δεν υπάρχει πιθανότητα για

την τιμή “false”. Εάν υποθέσουμε ότι η τιμή “false” αντιστοιχεί σε πιθανότητα 1, τότε η ψηφοφορία θα οδηγήσει σε μη ρεαλιστικά αποτελέσματα. Όλα τα σχήματα ψηφοφορίας που ορίστηκαν σε αυτή τη διατριβή αξιολογούνται συγκριτικά.

4.4 Διαφορές πιθανοτικής ψηφοφορίας με πολυστρατηγική μάθηση

Η χρήση πιθανοτικής ψηφοφορίας διαφέρει από τη χρήση πολυστρατηγικής μάθησης [48] ως προς τον τρόπο μοντελοποίησης του προβλήματος της εξαγωγής πληροφορίας σε μετα-επίπεδο. Η πολυστρατηγική μάθηση λογαριάζει κάθε πεδίο ξεχωριστά κατά το συνδυασμό και βασίζεται στον περιορισμό της ύπαρξης ενός μόνο παραδείγματος ανά κείμενο (OPD) για κάποια πεδία. Από την άλλη πλευρά, η πιθανοτική ψηφοφορία λαμβάνει χώρα πάνω στο συσσωρευμένο σχεδιότυπο, ενώ δεν υπάρχει κάποιος περιορισμός για τα σχετικά πεδία όπως στην πολυστρατηγική μάθηση. Κάτι τέτοιο επιτρέπει κατά το συνδυασμό την περίπτωση διφορούμενων πεδίων που έχουν προβλεφθεί από διαφορετικά συστήματα, όπως φαίνεται στον Πίνακα 4.2, όπου το τμήμα “1GB” έχει προβλεφθεί ως *ram* από το πρώτο σύστημα και ως *HDcapacity* από το δεύτερο. Το πεδίο με τη μεγαλύτερη πιθανότητα επιλέγεται στην περίπτωση κατά την ψηφοφορία. Αντίθετα, διφορούμενες προβλέψεις πεδίων αγνοούνται στην πολυστρατηγική μάθηση. Το Σχήμα 4.3 επιδεικνύει τη διαφορά μεταξύ πολυστρατηγικής μάθησης και πιθανοτικής ψηφοφορίας.



Μέθοδος Συνδυασμού	*Παραδείγματα που επιστρέφονται*
Πιθανοτική ψηφοφορία	<”256 MB”, ram> <”1 GB”, ram>
Πολυστρατηγική μάθηση	<”256 MB”, ram>

Σχήμα 4.3 Κατανόηση της διαφοράς μεταξύ ψηφοφορίας με πιθανοτήτες και πολυ-στρατηγικής μάθησης για το συνδυασμό δύο συστημάτων εξαγωγής πληροφορίας και όσον αφορά το πεδίο *ram*. Έστω $<"256\ MB", ram>$ το σωστό παράδειγμα.

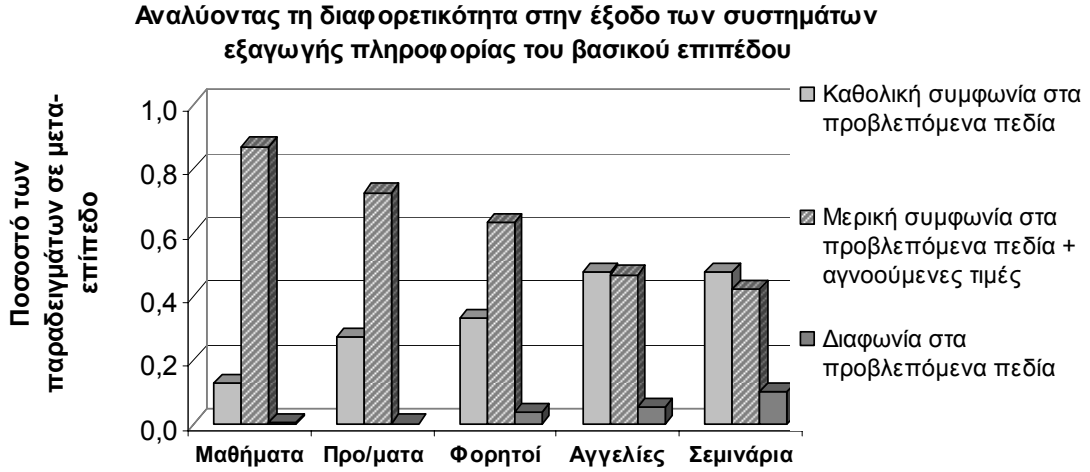
Σύμφωνα με το Σχήμα 4.3, συνδυάζοντας τα δύο συστήματα εξαγωγής πληροφορίας (που αναπαρίστανται με διαφορετικό τύπο γραμμής) μέσω πιθανοτικής ψηφοφορίας τότε επιστρέφονται και τα δύο τμήματα κειμένου που αναγνωρίζονται ως σχετικά για το πεδίο. Η πολυστρατηγική μάθηση εφαρμόζει τον περιορισμό OPD, επιλέγοντας το παράδειγμα $<"256\ MB", ram>$ ως σχετικό, το οποίο έχει τη μεγαλύτερη συνδυασμένη πιθανότητα και από τα δύο συστήματα. Στο Σχήμα 4.3 φαίνεται να πλεονεκτεί η πολυστρατηγική μάθηση έναντι της πιθανοτικής ψηφοφορίας που επιστρέφει ένα

επιπλέον παράδειγμα, το $<"1\,GB", ram>$, το οποίο είναι λανθασμένο. Εάν όμως το τελευταίο παράδειγμα ήταν σωστό, τότε η πολυστρατηγική μάθηση θα μειονεκτούσε, καθώς επιτρέπει μόνο ένα παράδειγμα σχετικού πεδίου ανά σελίδα κειμένου.

4.5 Ανάλυση των δεδομένων σε μετα-επίπεδο

Κάθε παράδειγμα σε μετα-επίπεδο αντιστοιχεί σε ένα τμήμα κειμένου $t(s,e)$ που έχει αναγνωριστεί ως σχετικό από τουλάχιστον ένα σύστημα εξαγωγής πληροφορίας, μαζί με τα πεδία που έχουν προβλεφθεί από τα συστήματα για το $t(s,e)$, τις αντίστοιχες πιθανότητες ορθότητας από τα συστήματα και τέλος το σωστό πεδίο για το $t(s,e)$ από τον ειδικό που έχει επισημειώσει τα κείμενα εκπαίδευσης.

Το Σχήμα 4.4 δείχνει έναν διαχωρισμό των παραδειγμάτων σε μετα-επίπεδο στο σύνολο των κειμένων επαλήθευσης για κάθε θεματική περιοχή, ανάλογα με το εάν όλα τα συστήματα του βασικού επιπέδου συμφωνούν στο ίδιο πεδίο για ένα τμήμα κειμένου $t(s,e)$ ή όχι. Ανάλογη ανάλυση της διαφορετικότητας στην έξοδο πολλαπλών συστημάτων εξαγωγής πληροφορίας απουσιάζει από τη διεθνή βιβλιογραφία.



Σχήμα 4.4 Διαχωρισμός των παραδειγμάτων του μετα-επιπέδου για κάθε θεματική περιοχή σε τρία διαφορετικά σύνολα, ανάλογα με το εάν όλα τα συστήματα συμφωνούν στο ίδιο πεδίο για ένα τμήμα κειμένου (αριστερή στήλη), ή κάποιο σύστημα - ή συστήματα - προβλέψει το ίδιο πεδίο ενώ τα υπόλοιπα δεν προβλέπουν τίποτα (μεσαία στήλη), ή υπάρχουν τουλάχιστον δύο διφορούμενες προβλέψεις πεδίων (δεξιά στήλη).

Η αριστερότερη στήλη για κάθε περιοχή στο Σχήμα 4.4 δείχνει ότι υπάρχουν χαρακτηριστικά των παραδειγμάτων σχετικών πεδίων στα κείμενα που μπορούν να αναγνωριστούν εύκολα από τα διαθέσιμα τρία συστήματα του βασικού επιπέδου, BWI,

HMMs και $(LP)^2$. Αφού ένα από τα συστήματα αυτά βασίζεται στη χρήση HMMs και χρησιμοποιεί μόνο λεκτικές μονάδες, δίχως κάποια άλλη γλωσσολογική πληροφορία, τότε συμπεραίνουμε ότι τα χαρακτηριστικά που αναγνωρίζουν και τα τρία συστήματα βασίζονται μόνο στις λεκτικές μονάδες. Για παράδειγμα, το "TFT" είναι ένα τυπικό παράδειγμα του πεδίου "τύπος οθόνης" για τους φορητούς υπολογιστές που εμφανίζεται συχνά στα κείμενα εκπαίδευσης και στα κείμενα επαλήθευσης και μπορεί να αναγνωριστεί εύκολα από όλα τα τρία διαθέσιμα συστήματα του βασικού επιπέδου.

Παρατηρώντας τη δεξιότερη στήλη για κάθε θεματική περιοχή στο Σχήμα 4.4 καταλήγουμε στο ιδιαίτερα ενδιαφέρον συμπέρασμα ότι είναι σπάνιες οι περιπτώσεις όπου τουλάχιστον δύο συστήματα εξαγωγής πληροφορίας του βασικού επιπέδου προβλέπουν διαφορετικά πεδία για ένα τμήμα κειμένου. Το τελευταίο συμπέρασμα μπορεί να εξηγηθεί εάν παρατηρήσουμε ότι τα συστήματα εξαγωγής πληροφορίας εκμεταλλεύονται τόσο το κείμενο που είναι σχετικό για την εξαγωγή *(target content)* όσο και το περιεχόμενο γύρω από αυτό *(surrounding context)* και επομένως μπορούν να ξεχωρίζουν παραδείγματα διαφορετικών πεδίων με παρόμοιο περιεχόμενο. Για παράδειγμα, παραδείγματα των πεδίων "ταχύτητα CD-rom" και "ταχύτητα DVD-rom" περιέχουν παρόμοιο περιεχόμενο, π.χ. "24x". Όμως τα συστήματα εξαγωγής πληροφορίας μπορούν κι εξετάζουν λεκτικές μονάδες γύρω από το "24x", όπως "cd" ή "dvd" κι επομένως μπορούν να διαχωρίσουν ανάμεσα στα δύο πεδία.

Μόνο σε μερικές περιπτώσεις αυτό δεν είναι εφικτό, οδηγώντας σε διφορούμενες προβλέψεις πεδίων, είτε διότι είναι δύσκολο να εντοπιστεί το περιεχόμενο εκείνο του κειμένου που βοηθά στον διαχωρισμό των παραδειγμάτων διαφορετικών πεδίων με παρόμοιο περιεχόμενο, είτε και λόγω περιορισμών στο περιεχόμενο κειμένου που εξερευνούν τα συστήματα του βασικού επιπέδου κατά την εκπαίδευσή τους. Για παράδειγμα, ο αλγόριθμος $(LP)^2$ εξερευνά ένα παράθυρο w λεκτικών μονάδων στα αριστερά και w λεκτικών μονάδων στα δεξιά των ορίων έναρξης και τέλους των παραδειγμάτων σχετικών πεδίων στα κείμενα εκπαίδευσης, όπου η τιμή για το w είναι προκαθορισμένη πριν την έναρξη της εκπαίδευσης. Στα HMMs, ένας προκαθορισμένος αριθμός *prefix* και *suffix* κόμβων μοντελοποιούν τις άμεσα προηγούμενες κι επόμενες λεκτικές μονάδες αντίστοιχα των επισημειωμένων παραδειγμάτων πεδίων στα κείμενα.

Για τις περιοχές των μαθημάτων της επιστήμης υπολογιστών και των ερευνητικών προγραμμάτων, οι διφορούμενες προβλέψεις πεδίων δεν υπερβαίνουν το 0.5 % όλων των δεδομένων σε μετα-επίπεδο. Για τους φορητούς υπολογιστές, το αντίστοιχο ποσοστό δεν υπερβαίνει το 3.5%. Για τις αγγελίες εργασίας, το ποσοστό είναι 5.5%,

ενώ τέλος, για τις ανακοινώσεις σεμιναρίων, το ποσοστό είναι 9.9%. Το ποσοστό είναι μεγαλύτερο για τους φορητούς υπολογιστές και τις αγγελίες εργασιών, εξαιτίας της ύπαρξης πολύ περισσοτέρων σχετικών πεδίων (19 και 17 αντίστοιχα), κάποια από τα οποία έχουν παρόμοιο περιεχόμενο. Στις ανακοινώσεις σεμιναρίων, η ώρα λήξης του σεμιναρίου (πεδίο *etime*) μπερδεύεται συχνά με την αντίστοιχη ώρα έναρξης (πεδίο *stime*), οδηγώντας σε διφορούμενες προβλέψεις για τα δύο πεδία, αυξάνοντας έτσι το μέγεθος της δεξιότερης στήλης στο Σχήμα 4.4 για τις ανακοινώσεις σεμιναρίων. Οι διφορούμενες προβλέψεις πεδίων προέρχονται κυρίως από το σύστημα των HMMs.

Αφού οι περιπτώσεις με διφορούμενες προβλέψεις πεδίων για ένα τμήμα κειμένου είναι αραιές, το ενδιαφέρον ερώτημα είναι το τι είδους διαφορετικότητα στην έξοδο των συστημάτων του βασικού επιπέδου μπορούν να εκμεταλλευτούν οι τεχνικές ψηφοφορίας και συσσωρευμένης γενίκευσης προκειμένου να επιτευχθούν καλύτερα αποτελέσματα εξαγωγής σε μετα-επίπεδο; Η απάντηση στο ερώτημα αυτό βρίσκεται στη μεσαία στήλη για κάθε θεματική περιοχή στο Σχήμα 4.4, η οποία δείχνει ότι η πλειοψηφία των παραδειγμάτων του μετα-επιπέδου αντιστοιχεί σε τμήματα κειμένου για τα οποία έχουν προβλεφθεί ίδια πεδία από μερικά αλλά όχι από όλα τα διαθέσιμα συστήματα εξαγωγής πληροφορίας, ενώ τα υπόλοιπα δεν προβλέπουν κάποιο πεδίο. Αφού χρησιμοποιούνται κατά την αξιολόγηση τρία συστήματα σε βασικό επίπεδο, αυτό αντιστοιχεί σε περιπτώσεις που είτε δύο συστήματα προβλέπουν το ίδιο πεδίο για ένα τμήμα κειμένου, ενώ το τρίτο δεν προβλέπει τίποτα, είτε μόνο ένα σύστημα προβλέπει κάποιο πεδίο με τα υπόλοιπα δύο συστήματα να μην προβλέπουν τίποτα.

Αναμένουμε, επομένως, από την ψηφοφορία και τη συσσωρευμένη γενίκευση να εξομαλύνουν αυτού του είδους τη διαφωνία, οδηγώντας σε καλύτερα αποτελέσματα εξαγωγής πληροφορίας σε μετα-επίπεδο. Θα ήταν επίσης ιδιαίτερα ενδιαφέρον να μελετηθεί η συμπεριφορά της συσσωρευμένης γενίκευσης όταν οι προβλέψεις από όλα τα συστήματα του βασικού επιπέδου συμφωνούν μεταξύ τους, σύμφωνα με την αριστερότερη στήλη για κάθε θεματική περιοχή στο Σχήμα 4.4.

Θα πρέπει να τονίσουμε ότι ο διαχωρισμός των δεδομένων του μετα-επιπέδου, όπως φαίνεται στο Σχήμα 4.4, πραγματοποιήθηκε για να επιτρέψει μια κατά το δυνατόν πιο αναλυτική και αντικειμενική σύγκριση των τεχνικών συνδυασμού που προτείνονται σε αυτή τη διατριβή, εξερευνώντας ταυτόχρονα τη συμπεριφορά τους βάσει του διαφορετικού βαθμού συσχέτισης στις προβλέψεις των συστημάτων του βασικού επιπέδου. Από την άλλη πλευρά ο Πίνακας 3.11 του Κεφαλαίου 3 δείχνει μια ποσοτική ανάλυση της διαφορετικότητας μεταξύ ζευγαριών συστημάτων του βασικού επιπέδου,

με στόχο να προσδιοριστεί εάν υπάρχει χώρος για περαιτέρω βελτίωση του καλύτερου συστήματος για κάθε θεματική περιοχή ενδιαφέροντος.

4.6 Αξιολόγηση τεχνικών ψηφοφορίας

Σε αυτή την ενότητα αξιολογούνται τα σχήματα ψηφοφορίας που περιγράφτηκαν στις Ενότητες 4.2 και 4.3, ενώ γίνεται σύγκριση και με την πολυστρατηγική μάθηση [48]. Η αξιολόγηση πραγματοποιείται για όλους τους δυνατούς συνδυασμούς των τριών συστημάτων εξαγωγής πληροφορίας που είναι διαθέσιμα σε βασικό επίπεδο. Τα αποτελέσματα που παρουσιάζονται στην ενότητα αυτή αποτελούν κατώφλι για την αξιολόγηση της συσσωρευμένης γενίκευσης για εξαγωγή πληροφορίας που ορίζεται και αξιολογείται στο επόμενο κεφάλαιο.

4.6.1 Αξιολόγηση πλειοψηφικής ψηφοφορίας

Έστω *MVotM* και *MVotF* οι δύο διαφορετικές περιπτώσεις πλειοψηφικής ψηφοφορίας, όπως ορίστηκαν στην Ενότητα 4.2. Στο *MVotM* οι αγνοούμενες τιμές πρόβλεψης από κάθε σύστημα εξαγωγής πληροφορίας παραβλέπονται, ενώ στο *MVotF* κωδικοποιούνται ως ειδικές ονομαστικές τιμές “false” υποδεικνύοντας άρνηση πρόβλεψης. Ο Πίνακας 4.3 δείχνει τα αποτελέσματα αξιολόγησης από τα *MVotM* και *MVotF*, μαζί με τα αποτελέσματα από το καλύτερο σύστημα εξαγωγής πληροφορίας του βασικού επιπέδου για κάθε θεματική περιοχή ενδιαφέροντος.

Πίνακας 4.3 Αποτελέσματα (%) των δύο διαφορετικών περιπτώσεων πλειοψηφικής ψηφοφορίας. *P = Ακρίβεια (Precision)*, *R = Ανάκληση (Recall)*.

	Καλύτερο σύστημα του βασικού επιπέδου			Πλειοψηφική Ψηφοφορία (*MVotM)*			Πλειοψηφική Ψηφοφορία (*MVotF)*		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Μαθήματα	71.39	60.90	65.73	58.68	74.35	65.59	82.05	47.65	60.29
Προγράμματα	56.24	68.18	61.64	49.17	79.32	60.71	68.88	65.96	67.39
Φορητοί	62.29	65.42	63.81	52.89	76.00	62.37	80.41	59.05	68.09
Αγγελίες	87.70	79.18	83.22	71.29	90.88	79.90	93.06	76.31	83.85
Σεμινάρια	91.39	81.63	86.23	86.93	86.82	86.87	97.55	78.72	87.13

Μια σημαντική παρατήρηση είναι ότι το *MVotM* βελτιώνει την ανάκληση αλλά βλάπτει την ακρίβεια, σε σύγκριση με το καλύτερο σύστημα του βασικού επιπέδου για κάθε θεματική περιοχή. Αντίθετα, το *MVotF* βελτιώνει την ακρίβεια αλλά βλάπτει την ανάκληση. Αυτή η αντικρουόμενη συμπεριφορά από τα δύο διαφορετικά σχήματα πλειοψηφικής ψηφοφορίας οφείλεται κυρίως σε περιπτώσεις όπου μόνο ένα σύστημα προβλέπει ένα πεδίο f για ένα τμήμα κειμένου, ενώ τα υπόλοιπα δύο συστήματα δεν

προβλέπουν κάποιο πεδίο. Στις περιπτώσεις αυτές, εάν f δεν είναι το σωστό πεδίο για το τμήμα κειμένου, βλάπτεται η ακρίβεια για το *MVotM* που πάντα επιστρέφει το f. Αντιθέτως, εάν το f είναι το σωστό πεδίο, βλάπτεται η ανάκληση για το *MVotF* αφού θα επιστρέψει την τιμή “false”, αφού είναι εκείνη με το μεγαλύτερο αριθμό ψήφων.

Στον Πίνακα 4.3 φαίνεται επίσης ότι η ανάκληση βλάπτεται περισσότερο από το *MVotF* στη θεματική περιοχή των μαθημάτων της επιστήμης υπολογιστών, με συνεπακόλουθο να βλάπτεται περισσότερο και η μετρική $F1$. Κάτι τέτοιο σημαίνει ότι οι “μονές” προβλέψεις, δηλαδή όταν μόνο ένα από τα τρία διαθέσιμα συστήματα προβλέπει κάποιο πεδίο για ένα τμήμα κειμένου, είναι τις περισσότερες φορές λανθασμένες για την περιοχή αυτή. Στις περιπτώσεις ισοπαλίας στις προβλέψεις πεδίων για ένα τμήμα κειμένου, τότε τυχαία επιλογή πραγματοποιείται ανάμεσα στις ισόπαλες τιμές. Βέβαια περιπτώσεις διφορούμενων πεδίων για τμήματα κειμένου συμβαίνουν σε μικρό ποσοστό στις θεματικές περιοχές που εξετάζονται σε αυτή τη διατριβή, όπως φαίνεται και από τη δεξιά στήλη για κάθε περιοχή στο Σχήμα 4.4.

Μια σημαντική παρατήρηση επίσης είναι ότι αφού σπάνια συμβαίνουν διφορούμενες προβλέψεις πεδίων στις θεματικές περιοχές που εξετάζονται, τότε το *MVotM* δεν πραγματοποιεί ουσιαστική ψηφοφορία. Όπως εξάλλου φαίνεται και από την αριστερή και τη μεσαία στήλη για κάθε περιοχή στο Σχήμα 4.4, στη συντριπτική πλειοψηφία των δεδομένων του μετα-επιπέδου, μόνο ένα πεδίο f συμμετέχει κατά τη διαδικασία της πλειοψηφικής ψηφοφορίας. Όσο περισσότερα συστήματα του βασικού επιπέδου προβλέψουν το ίδιο πεδίο f για ένα τμήμα κειμένου $t(s,e)$, τόσο μεγαλύτερος ο τελικός αριθμός ψήφων για το f που τελικά επιστρέφεται ως πρόβλεψη για το $t(s,e)$. Εάν όμως το f δεν είναι η σωστή πρόβλεψη για το $t(s,e)$, τότε δεν υπάρχει τρόπος να το απορρίψουμε, αφού αγνοούμενες προβλέψεις παραβλέπονται κατά το μέτρημα των ψήφων, βλάπτοντας έτσι την ακρίβεια στην εξαγωγή.

Ως αποτέλεσμα της παραπάνω παρατήρησης, η ανάκληση που επιτυγχάνεται από το *MVotM* είναι μια αρκετά καλή προσέγγιση της μέγιστης ανάκλησης που μπορεί να επιτευχθεί στην εξαγωγή πληροφορίας σε μετα-επίπεδο για κάθε θεματική περιοχή. Αυτό συμβαίνει γιατί κάθε σύστημα του βασικού επιπέδου συνεισφέρει στο συνδυασμό τα σωστά παραδείγματα πεδίων που μοναδικά αυτό αναγνωρίζει. Αντίστοιχα, τα λανθασμένα παραδείγματα πεδίων που μοναδικά προβλέπει κάθε σύστημα βλάπτουν την ακρίβεια. Από την άλλη πλευρά, το *MVotF* συμπεριφέρεται περισσότερο ως ψηφοφορία, αφού κάθε αγνοούμενη πρόβλεψη κωδικοποιείται ως μια ειδική τιμή “false” η οποία συμμετέχει κατά τη διαδικασία μέτρησης των ψήφων.

Το τελικό συμπέρασμα είναι ότι κανένα από τα δύο διαφορετικά σχήματα πλειοψηφικής ψηφοφορίας δεν αποδείχτηκε καθολικά αποτελεσματικό, με βάση τη μετρική $F1$, και στις πέντε θεματικές περιοχές. Το *MVotF* επιτυγχάνει μεγαλύτερο και ταυτόχρονα στατιστικά σημαντικότερο $F1$ σε μετα-επίπεδο για τις θεματικές περιοχές των ερευνητικών προγραμμάτων και ανακοινώσεων σεμιναρίων, ενώ το *MVotM* δεν βελτιώνει το καλύτερο $F1$ του βασικού επιπέδου σε καμία από τις πέντε περιοχές. Η μεγάλη βελτίωση στο $F1$ που επιτυγχάνει το *MVotF* στους φορητούς υπολογιστές δεν είναι συνεπής κατά την αξιολόγηση σε όλα τα βήματα της διασταυρωμένης επικύρωσης που ακολουθήθηκε, και γι' αυτό μετρήθηκε ως στατιστικά μη σημαντική. Επιπρόσθετα, θα ήταν επιθυμητό να *μαθαίναμε* για το εάν μια πρόβλεψη πεδίου f για ένα τμήμα κειμένου είναι σωστή, αντί να δεχόμαστε πάντα την πρόβλεψη αυτή, όπως στην περίπτωση του *MVotM*, εάν το f έχει το μεγαλύτερο αριθμό ψήφων, ή να απορρίπτουμε πάντα το f, εάν η τιμή με τον μεγαλύτερο αριθμό ψήφων είναι η τιμή "false". Αυτός είναι ένας σημαντικός στόχος για τη συσσωρευμένη γενίκευση.

4.6.2 Αξιολόγηση πιθανοτικής ψηφοφορίας

Έστω *PVotM* και *PVotF* οι δύο διαφορετικές περιπτώσεις ψηφοφορίας με χρήση πιθανοτήτων, όπως ορίστηκαν στην Ενότητα 4.3. Στο *PVotM*, τυχόν αγνοούμενες προβλέψεις για ένα τμήμα κειμένου παραβλέπονται, όπως ακριβώς και στην πρώτη περίπτωση της πλειοψηφικής ψηφοφορίας *(MVotM)*. Στο *PVotF*, εάν η υψηλότερη (συνδυασμένη) πιθανότητα για ένα πεδίο f είναι μικρότερη από 0.5, τότε η πρόβλεψη για το f απορρίπτεται. Ο Πίνακας 4.4 δείχνει τα αποτελέσματα που επιτευχθήκαν από τα *PVotM* και *PVotF* για κάθε θεματική περιοχή, μαζί με τα καλύτερα αποτελέσματα που επιτεύχθηκαν σε βασικό επίπεδο για κάθε περιοχή.

Πίνακας 4.4 Αποτελέσματα (%) των δύο διαφορετικών περιπτώσεων πιθανοτικής ψηφοφορίας. *P = Ακρίβεια (Precision)*, *R = Ανάκληση (Recall)*.

	Καλύτερο σύστημα του βασικού επιπέδου			Πιθανοτική ψηφοφορία (*PVotM*)			Πιθανοτική ψηφοφορία (*PVotF*)		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Μαθήματα	71.39	60.90	65.73	58.78	74.35	65.65	70.16	71.12	70.64
Προγράμματα	56.24	68.18	61.64	49.20	79.37	60.75	63.31	68.38	65.75
Φορητοί	62.29	65.42	63.81	53.21	76.47	62.76	72.86	69.30	71.03
Αγγελίες	87.70	79.18	83.22	71.37	90.98	79.99	80.08	86.45	83.15
Σεμινάρια	91.39	81.63	86.23	86.99	86.82	86.90	90.69	85.50	88.02

Η συμπεριφορά του *PVotM* είναι παρόμοια με εκείνη του *MVotM*, όπως φαίνεται και από τη σύγκριση των Πινάκων 4.3 και 4.4. Αφού η συντριπτική πλειοψηφία των παραδειγμάτων σε μετα-επίπεδο δεν περιλαμβάνει διφορούμενες προβλέψεις πεδίων, η χρήση πιθανοτήτων δεν αποδεικνύεται ιδιαίτερα ωφέλιμη για το *PVotM*.
Όσο μεγαλύτερος είναι ο αριθμός των ψήφων για ένα πεδίο f, τόσο μεγαλύτερη είναι και η συνδυασμένη πιθανότητα ορθότητας για το πεδίο αυτό. Μόνο για τους φορητούς ηλεκτρονικούς υπολογιστές, η πολύ μικρή βελτίωση στο $F1$ που επιτυγχάνει το *PVotM* έναντι του *MVotM*, μετρήθηκε ως στατιστικά σημαντική.

Σε αντίθεση με την πλειοψηφική ψηφοφορία και το *PVotM*, το *PVotF* επιτυγχάνει συγκρίσιμη (μόνο για τις αγγελίες εργασίας) ή καλύτερη απόδοση στην εξαγωγή πληροφορίας σε όλες τις θεματικές περιοχές, σε σύγκριση με το καλύτερο σύστημα σε βασικό επίπεδο. Συγκρίνοντας όλα τα διαφορετικά σχήματα ψηφοφορίας μεταξύ τους, το *PVotF* είναι το καλύτερο για τις θεματικές περιοχές των πανεπιστημιακών μαθημάτων της επιστήμης υπολογιστών, των φορητών ηλεκτρονικών υπολογιστών και για τις ανακοινώσεις σεμιναρίων. Το *MVotF* συμπεριφέρεται καλύτερα μόνο για τις αγγελίες εργασίας. Στην περιοχή των ερευνητικών προγραμμάτων, η βελτίωση στο $F1$ που επιτυγχάνει το *MVotF* σε σχέση με το *PVotF* μετρήθηκε ως στατιστικά μη σημαντική.

Το *PVotF* πραγματοποιεί έναν επιπλέον έλεγχο στο πεδίο f που επιστρέφεται από το *PVotM*, εξετάζοντας εάν η πιθανότητα που υπολογίζεται για το f είναι μεγαλύτερη ή όχι από την τιμή 0.5 και αναλόγως αποδέχεται ή απορρίπτει το f. Αυτή η τακτική οδηγεί σε πιο ακριβή αποτελέσματα για το *PVotF*, σε σύγκριση με το *PVotM*, αλλά και σε χειρότερη ανάκληση. Η βελτίωση όμως στην ακρίβεια είναι μεγαλύτερη από ότι η απώλεια στην ανάκληση, οδηγώντας σε μεγαλύτερο $F1$ για το *PVotF*, αφού οι περισσότερες λανθασμένες προβλέψεις έχουν πιθανότητα μικρότερη από 0.5.

Επιπρόσθετα, εάν η τιμή με τον μεγαλύτερο αριθμό ψήφων είναι η τιμή “false” (δηλαδή δύο από τα τρία συστήματα του βασικού επιπέδου απέχουν από την πρόβλεψη), το *PVotF* εξετάζει αντιστοίχως την πιθανότητα το πεδίου f με τον επόμενο μεγαλύτερο αριθμό ψήφων (δηλαδή της πρόβλεψης του εναπομείναντος τρίτου συστήματος). Αυτό εξηγεί επίσης τη υψηλότερη ανάκληση και ταυτόχρονα τη χαμηλότερη ακρίβεια που επιτυγχάνει το *PVotF*, σε σχέση με το *MVotF* που πάντα επιστρέφει “false” στην περίπτωση αυτή. Για τις περιοχές των ερευνητικών προγραμμάτων και αγγελιών εργασίας, όμως, η μείωση στην ακρίβεια από το *PVotF* σε σχέση με το *MVotF* είναι μεγαλύτερη από ότι το όφελος στην ανάκληση, οδηγώντας έτσι σε συγκρίσιμο η ανώτερο $F1$ για το *MVotF*. Αυτό οφείλεται στις περισσότερες περιπτώσεις, για τις δύο

αυτές περιοχές, όπου μονές προβλέψεις (δηλαδή προβλέψεις μόνο ενός από τα τρία συστήματα) είναι ταυτόχρονα και λανθασμένες και έχουν και μεγάλη πιθανότητα, και για το λόγο αυτό ορθά απορρίπτονται από το *MVotF* ενώ λανθασμένα γίνονται δεκτές από το σχήμα ψηφοφορίας *PVotF*.

Οι Πίνακες 4.5 έως 4.7 συνοψίζουν τη σύγκριση ανάμεσα σε όλα τα σχήματα ψηφοφορίας και στο καλύτερο σύστημα του βασικού επιπέδου, με βάση τον αριθμό των στατιστικά πιο σημαντικών *νικών* έναντι *ηττών* στις πέντε θεματικές περιοχές, χρησιμοποιώντας τις τρεις μετρικές αξιολόγησης (ακρίβεια, ανάκληση, $F1$) αντίστοιχα.

Πίνακας 4.5 Στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, χρησιμοποιώντας τη μετρική *ακρίβεια*, στις πέντε θεματικές περιοχές ενδιαφέροντος.

	Βασικό	MVotM	MVotF	PVotM	PVotF
Βασικό		5\0	0\5	5\0	2\2
MVotM	0\5		0\5	0\1	0\5
MVotF	5\0	5\0		5\0	5\0
PVotM	0\5	1\0	0\5		0\5
PVotF	2\2	5\0	0\5	5\0	

Πίνακας 4.6 Στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, χρησιμοποιώντας τη μετρική *ανάκληση*, στις πέντε θεματικές περιοχές ενδιαφέροντος.

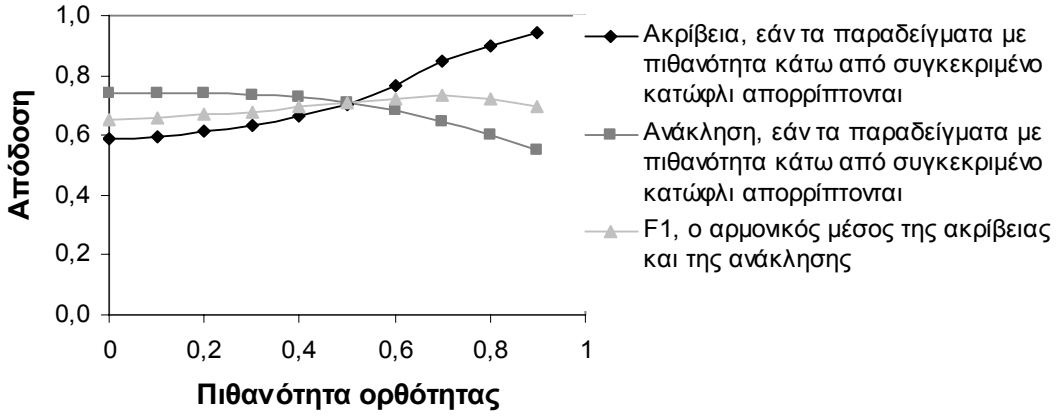
	Βασικό	MVotM	MVotF	PVotM	PVotF
Βασικό		0\5	4\0	0\5	0\4
MVotM	5\0		5\0	0\1	5\0
MVotF	0\4	0\5		0\5	0\5
PVotM	5\0	1\0	5\0		5\0
PVotF	4\0	0\5	5\0	0\5	

Πίνακας 4.7 Στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, χρησιμοποιώντας τη μετρική $F1$, στις πέντε θεματικές περιοχές ενδιαφέροντος.

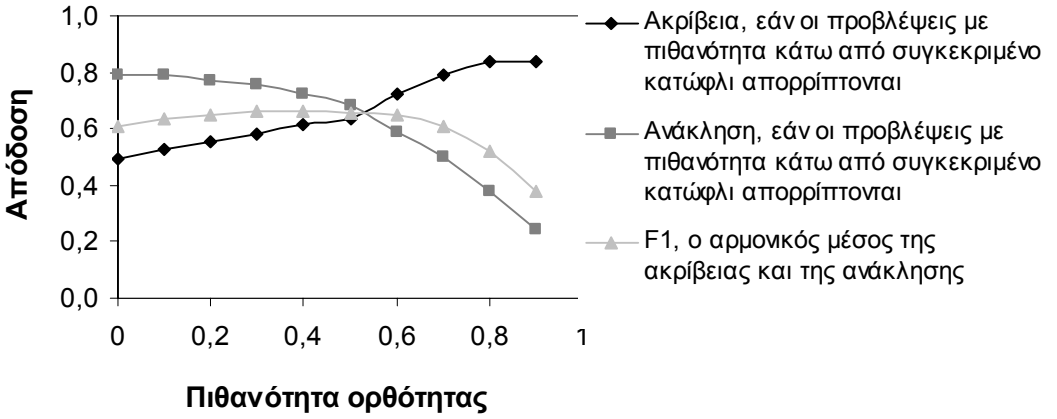
	Βασικό	MVotM	MVotF	PVotM	PVotF
Βασικό		2\0	1\2	1\0	0\4
MVotM	0\2		1\3	0\1	0\5
MVotF	2\1	3\1		3\1	1\3
PVotM	0\1	1\0	1\3		0\5
PVotF	4\0	5\0	3\1	5\0	

4.6.3 Διαχείριση του κατωφλίου αποδοχής/απόρριψης προβλέψεων

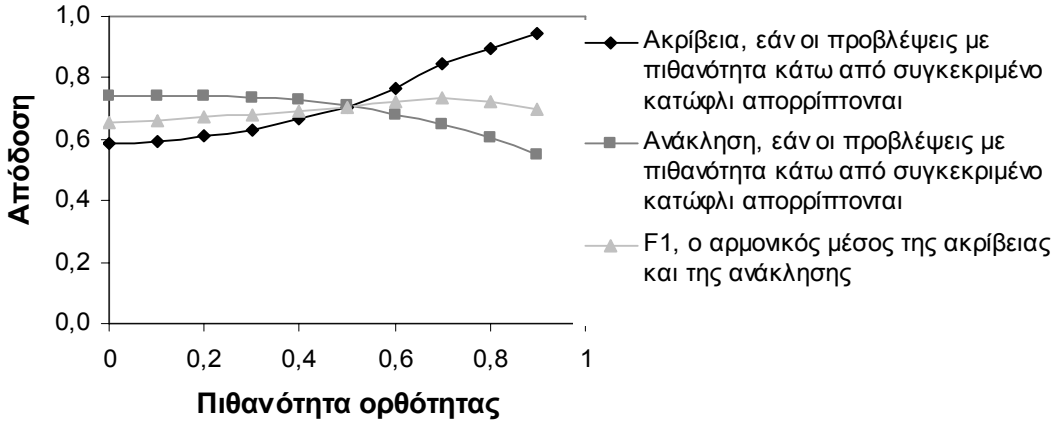
Παρόλο που η χρήση πιθανοτήτων ορθότητας είναι καλύτερη από τη χρήση απλών βαθμών εμπιστοσύνης στην έξοδο των συστημάτων του βασικού επιπέδου, επιλέγοντας ως κατώφλι την τιμή 0.5, για την αποδοχή/απόρριψη, μπορεί να μην είναι πάντα η βέλτιστη επιλογή, όπως φαίνεται στα Σχήματα 4.5 έως 4.9 για τις πέντε θεματικές περιοχές ενδιαφέροντος.



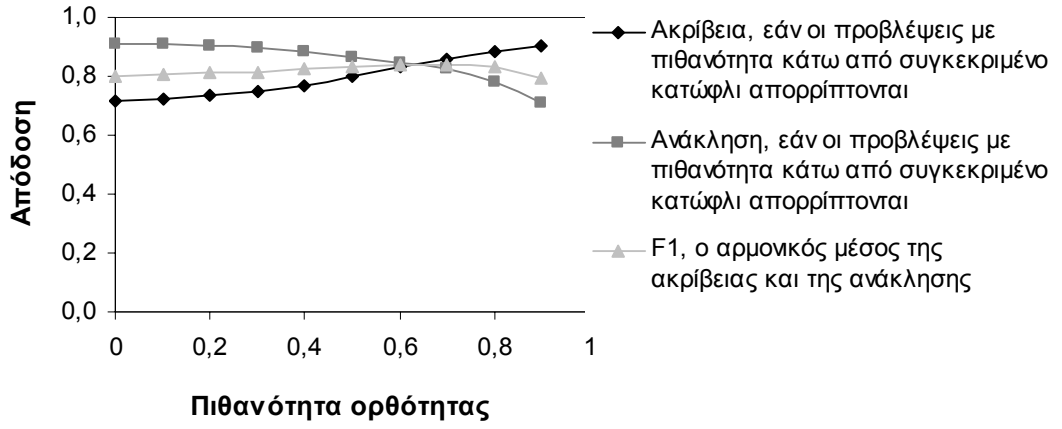
Σχήμα 4.5 Ανάκληση, ακρίβεια και *F1*, από το *PVotF* για τη θεματική περιοχή των μαθημάτων επιστήμης υπολογιστών, σε σχέση με το κατώφλι απόρριψης.



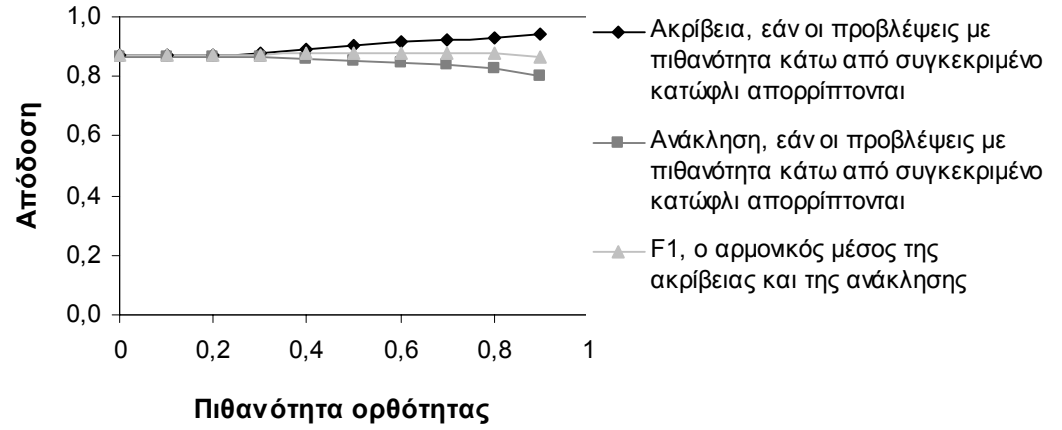
Σχήμα 4.6 Ανάκληση, ακρίβεια και *F1*, από το *PVotF* για τη θεματική περιοχή των ερευνητικών προγραμμάτων, σε σχέση με το κατώφλι απόρριψης.



Σχήμα 4.7 Ανάκληση, ακρίβεια και $F1$, από το *PVotF* για τη θεματική περιοχή των φορητών ηλεκτρονικών υπολογιστών, σε σχέση με το κατώφλι απόρριψης.



Σχήμα 4.8 Ανάκληση, ακρίβεια και $F1$, από το *PVotF* για τη θεματική περιοχή των αγγελιών εργασίας, σε σχέση με το κατώφλι απόρριψης.



Σχήμα 4.9 Ανάκληση, ακρίβεια και $F1$, από το PVotF για τη θεματική περιοχή των ανακοινώσεων σεμιναρίων, σε σχέση με το κατώφλι απόρριψης.

Η απόδοση του *PVotF* (ανάκληση, ακρίβεια και $F1$) για κατώφλι απόρριψης μηδέν, ταυτίζεται με εκείνη του *PVotM*. Αυξάνοντας το κατώφλι, κάτω από το οποίο προβλέψεις πεδίων απορρίπτονται, μια ανταλλαγή στην απόδοση παρατηρείται μεταξύ της ακρίβειας και της ανάκλησης, αντανακλώντας το φυσικό γεγονός ότι μια πρόβλεψη που απορρίπτεται μπορεί να είναι είτε σωστή, βλάπτοντας έτσι και την ακρίβεια και την ανάκληση, είτε λανθασμένη, βελτιώνοντας κατά συνέπεια μόνο την ακρίβεια.

Όπως φαίνεται και στα Σχήματα 4.5 έως 4.9, οι περισσότερες λανθασμένες προβλέψεις απορρίπτονται όσο αυξάνεται το κατώφλι απόρριψης, αφού οι περισσότερες ορθές προβλέψεις πεδίων έχουν μεγάλη πιθανότητα ορθότητας.

Η μετατροπή βαθμών εμπιστοσύνης σε πιθανότητες ορθότητας στις προβλέψεις των συστημάτων του βασικού επιπέδου, γίνεται με αξιολόγηση της απόδοσής τους σε ένα

ξεχωριστό σύνολο κειμένων, ή με διασταυρωμένη επικύρωση στα κείμενα εκπαίδευσης. Αυτό μονάχα προσεγγίζει την *αληθινή* πιθανοτική κατανομή ορθότητας στις προβλέψεις κάθε συστήματος στο σύνολο όλων των κειμένων μιας θεματικής περιοχής που μπορούν να απαριθμηθούν. Ως αποτέλεσμα, η βέλτιστη επιλογή του κατωφλίου απόρριψης προβλέψεων πεδίων μπορεί να διαφέρει για διαφορετικές συλλογές κειμένων και να μην είναι 0.5. Στο Σχήμα 4.5 για παράδειγμα, παρόλο που η ακρίβεια και η ανάκληση αποκτούν την ίδια τιμή για κατώφλι 0.5, η καλύτερη τιμή για το $F1$ (73%) επιτυγχάνεται για κατώφλι 0.7. Το καλύτερο $F1$ παραμένει στο κατώφλι 0.5 για τα ερευνητικά προγράμματα και τις ανακοινώσεις σεμιναρίων, ενώ για τους φορητούς υπολογιστές, το καλύτερο $F1$ (71.2%) επιτυγχάνεται για κατώφλι 0.6. Τέλος, για τις αγγελίες εργασίας, το βέλτιστο $F1$ (84.08%) είναι στο κατώφλι 0.7.

Το τελικό συμπέρασμα είναι ότι η ψηφοφορία με πιθανότητες, θέτοντας ένα κατώφλι αποδοχής/απόρριψης στις προβλέψεις πεδίων, αν και είναι η καλύτερη μέθοδος ψηφοφορίας, δεν είναι καθολικά αποτελεσματική και στις πέντε θεματικές περιοχές ενδιαφέροντος. Ο Πίνακας 4.4 δείχνει ότι στην περιοχή των μαθημάτων, το σχήμα *PVotF* δε βελτιώνει το καλύτερο $F1$ του βασικού επιπέδου. Επιπλέον, αντί να προσπαθούμε να βελτιστοποιούμε την επιλογή κατωφλίου με έναν τέτοιο εμπειρικό τρόπο, θα ήταν πιο ενδιαφέρον να προσπαθήσουμε να *μάθουμε* πότε μια πρόβλεψη πεδίου είναι σωστή. Αυτός θα είναι επίσης ένας σημαντικός στόχος για τη συσσωρευμένη γενίκευση.

4.6.4 Αξιολόγηση πολυστρατηγικής μάθησης

Ο Πίνακας 4.8 συγκρίνει τις τιμές για το $F1$ που επιτεύχθηκαν από την πολυστρατηγική μάθηση, όπως έχει περιγραφεί στην Ενότητα 2.5, με τις καλύτερες αντίστοιχες τιμές του βασικού επιπέδου, την ψηφοφορία με χρήση πιθανοτήτων και τα αποτελέσματα της πολυστρατηγικής μάθησης, όπως παρουσιάζονται από τον Freitag [48]. Η σύγκριση πραγματοποιείται για τις θεματικές περιοχές των μαθημάτων της επιστήμης υπολογιστών, των ερευνητικών προγραμμάτων και των ανακοινώσεων σεμιναρίων, αφού για τις περιοχές αυτές παρουσιάζει αποτελέσματα και ο Freitag [48].

Τα αποτελέσματα από την πολυστρατηγική μάθηση είναι καλύτερα από τα αντίστοιχα του Freitag [48] και οφείλεται μερικώς στην καλύτερη απόδοση των τριών συστημάτων, BWI, HMMs και $(LP)^2$, που χρησιμοποιούνται σε βασικό επίπεδο στη διατριβή αυτή.

Πίνακας 4.8 Καλύτερο $F1$ (%) ανά σχετικό πεδίο για τις περιοχές των μαθημάτων επιστήμης υπολογιστών, ερευνητικών προγραμμάτων και ανακοινώσεων σεμιναρίων, που επιτυγχάνονται από το καλύτερο σύστημα του βασικού επιπέδου, την πιθανοτική ψηφοφορία και την πολυστρατηγική μάθηση.

	Καλύτερο βασικό	Ψηφοφορία PVotM	Ψηφοφορία PVotF	Πολύ-στρατηγική μάθηση	Πολύ-στρατ/κή Μάθηση [48]
crsNumber	94.46	95.72	95.92	94.91	88.9
crsTitle	70.05	73.50	72.34	73.29	62.0
crsInst	48.21	50.81	57.76	50.81	49.8
projMember	65.00	63.16	68.94	63.09	45.5
projTitle	39.66	34.26	32.88	35.65	34.1
stime	99.09	99.51	99.51	99.42	99.3
etime	97.62	89.09	96.68	67.15	94.3
speaker	73.41	75.88	75.40	75.58	66.2
location	77.43	81.82	81.83	81.82	79.7

Η ομοιότητα ανάμεσα στην πολυστρατηγική μάθηση και στην πιθανοτική ψηφοφορία που παραβλέπει αγνοούμενες τιμές (*PVotM*), όπως έχει περιγραφτεί στην Ενότητα 4.3, εξηγεί το γεγονός γιατί οι δύο μέθοδοι συμπεριφέρονται διαφορετικά για το πεδίο *etime* και όμοια για τα υπόλοιπα πεδία. Παραδείγματα του πεδίου *etime* συγχέονται αρκετές φορές με παραδείγματα του πεδίου *stime*, από τα συστήματα του βασικού επιπέδου, οδηγώντας έτσι σε διφορούμενες προβλέψεις για τα δύο αυτά πεδία. Δηλαδή τμήματα κειμένου που αντιστοιχούν σε ώρες έναρξης σεμιναρίου (*stime*) αναγνωρίζονται *και* ως παραδείγματα ώρας λήξης σεμιναρίου (*etime*).

Η πολυστρατηγική μάθηση, όμως, αντιμετωπίζει κάθε πεδίο ξεχωριστά κατά το συνδυασμό, αποτυγχάνοντας να διαχωρίσει ανάμεσα στα δύο αυτά πεδία. Η πιθανοτική ψηφοφορία (ειδικότερα η τεχνική *PVotF*), από την άλλη πλευρά, αντιμετωπίζει καλύτερα τις διφορούμενες προβλέψεις πεδίων, επιλέγοντας εκείνο με τη μεγαλύτερη (συνδυασμένη) πιθανότητα, οδηγώντας σε πολύ καλύτερα αποτελέσματα έναντι της πολυστρατηγικής μάθησης για το *etime*. Το γεγονός ότι δε βελτιώνεται από το *PVotF* το καλύτερο $F1$ του βασικού επιπέδου για το *etime*, δεν οφείλεται σε λανθασμένη επιλογή έναντι του *stime* (γιατί έτσι θα είχαμε χειρότερα αποτελέσματα και για το *stime*, κάτι που δεν επιβεβαιώνεται από τον Πίνακα 4.8), αλλά σε κάποιες περιπτώσεις που τμήματα κειμένου έχουν λανθασμένα αναγνωριστεί ως *etime* με πιθανότητα μεγαλύτερη από 0.5 και άρα δεν απορρίπτονται από το *PVotF*.

Διφορούμενες προβλέψεις πεδίων από τα συστήματα του βασικού επιπέδου, συμβαίνουν λιγότερο συχνά για τα υπόλοιπα πεδία στις πέντε περιοχές, όπως φαίνεται και στο Σχήμα 4.4. Επομένως το *PVotM*, αλλά και το *PVotF*, χειρίζεται κάθε πεδίο ξεχωριστά κατά το συνδυασμό, όπως κάνει εξ ορισμού η πολυστρατηγική μάθηση.

Όπως και στο *PVotM*, η επιτυχία της πολυστρατηγικής μάθησης εξαρτάται ισχυρά από το πώς οι λανθασμένες προβλέψεις από όλα τα συστήματα του βασικού επιπέδου συσχετίζονται μεταξύ τους. Όσο περισσότερα λανθασμένα παραδείγματα πεδίων αναγνωρίζει μοναδικά κάθε σύστημα, τόσο μεγαλύτερη είναι η ζημιά στην ακρίβεια, αφού τυχόν αγνοούμενες προβλέψεις παραβλέπονται και δεν υπάρχει σωστό πεδίο να αντικρούσει το λανθασμένο κατά την ψηφοφορία. Αυτό μπορεί να οδηγήσει σε χειρότερο $F1$, σε σχέση με το καλύτερο σύστημα του βασικού επιπέδου, όπως συμβαίνει στα ερευνητικά προγράμματα.

Το τελικό συμπέρασμα είναι ότι το σχήμα *PVotF* είναι το καλύτερο απ' όσα σχήματα ψηφοφορίας περιγράφονται σε αυτό το κεφάλαιο, σύμφωνα με την αξιολόγηση στις πέντε θεματικές περιοχές ενδιαφέροντος και χρησιμοποιώντας τα τρία συστήματα εξαγωγής πληροφορίας, BWI, HMMs και $(LP)^2$, που είναι διαθέσιμα σε βασικό επίπεδο.

4.6.5 Αξιολόγηση ψηφοφορίας σε ζευγάρια συστημάτων

Πειράματα πραγματοποιήθηκαν επίσης και σε συνδυασμούς ζευγαριών συστημάτων εξαγωγής πληροφορίας. Ο στόχος ήταν να διαπιστωθεί εάν η ψηφοφορία των τριών συστημάτων του βασικού επιπέδου οδηγεί σε καλύτερα αποτελέσματα εξαγωγής σε μετα-επίπεδο, από ότι ο συνδυασμός ζευγαριών συστημάτων. Το Παράρτημα Α.7 δείχνει αναλυτικά τα αποτελέσματα που επιτεύχθηκαν, σε μετα-επίπεδο στις πέντε περιοχές από τις τεχνικές ψηφοφορίας που περιγράφονται σε αυτό το κεφάλαιο, για όλους τους δυνατούς συνδυασμούς συστημάτων και μετρικών αξιολόγησης.

Ο Πίνακας 4.9 συνοψίζει τις καλύτερες τιμές για το $F1$ που επιτευχθήκαν κατά την ψηφοφορία για όλους τους δυνατούς συνδυασμούς συστημάτων σε βασικό επίπεδο, και δείχνει ότι τις περισσότερες φορές το σχήμα *PVotF* οδηγεί στο καλύτερο $F1$. Ο ίδιος πίνακας δείχνει επίσης ότι ορισμένοι συνδυασμοί ζευγαριών συστημάτων επιτυγχάνουν συγκρίσιμα αποτελέσματα σε σχέση με το συνδυασμό και των τριών συστημάτων.

Πίνακας 4.9 Καλύτερες τιμές για το $F1$ (%) από τις τεχνικές ψηφοφορίας για όλους τους συνδυασμούς των τριών συστημάτων εξαγωγής πληροφορίας σε βασικό επίπεδο.

	BWI+HMMs		BWI+$(LP)^2$		HMMs+$(LP)^2$		BWI+HMMs+$(LP)^2$	
Μαθήματα	60.25	*MVotM*	69.46	*PVotF*	71.60	*PVotF*	70.64	*PVotF*
Προγράμματα	61.60	*MVotF*	63.38	*MVotM*	63.97	*PVotF*	67.39	*MVotF*
Φορητοί	66.95	*PVotF*	67.23	*PVotF*	71.23	*PVotF*	71.03	*PVotF*
Αγγελίες	80.14	*PVotF*	83.86	*MVotF*	82.44	*PVotF*	83.85	*MVotF*
Σεμινάρια	86.57	*PVotF*	87.12	*PVotF*	87.85	*PVotF*	88.02	*PVotF*

Κάτι τέτοιο συμβαίνει στο συνδυασμό, μέσω *PVotF*, των HMMs με $(LP)^2$ καθώς και σε εκείνο του BWI με $(LP)^2$ στην περιοχή των μαθημάτων της επιστήμης υπολογιστών, στο συνδυασμό των HMMs με $(LP)^2$ στους φορητούς υπολογιστές και στις ανακοινώσεις σεμιναρίων, και τέλος στο συνδυασμό BWI με $(LP)^2$, αλλά μέσω *MVotF*, στην περιοχή των αγγελιών εργασίας. Κανένας όμως συνδυασμός ζευγαριών συστημάτων δεν επιτυγχάνει στατιστικά καλύτερη τιμή για το $F1$, σε σχέση με το συνδυασμό και των τριών συστημάτων, σε καμία θεματική περιοχή ενδιαφέροντος.

Το Παράρτημα Α.7 δείχνει επίσης ότι ο συνδυασμός περισσοτέρων συστημάτων, είτε μέσω *MVotM* είτε μέσω *PVotM*, οδηγεί σε μεγαλύτερη ανάκληση, λόγω των σωστών παραδειγμάτων που μοναδικά αναγνωρίζει κάθε σύστημα. Η ακρίβεια όμως βλάπτεται, λόγω των λανθασμένων παραδειγμάτων που μοναδικά επίσης αναγνωρίζει κάθε σύστημα. Μόνο στα μαθήματα της επιστήμης υπολογιστών, παρατηρούμε ότι προσθέτοντας το σύστημα του $(LP)^2$ στο συνδυασμό, είτε μέσω *MVotM* είτε μέσω *PVotM*, των άλλων δύο συστημάτων, οδηγεί σε ελαφρά αύξηση της ακρίβειας σε μετα-επίπεδο. Αυτό οφείλεται στη μεγαλύτερη συσχέτιση, για την περιοχή αυτή, των λανθασμένων παραδειγμάτων ανάμεσα στο σύστημα του $(LP)^2$ και τα άλλα δύο συστήματα, ταυτόχρονα με τη μικρή συσχέτιση στα σωστά παραδείγματα.

Μεγάλη συσχέτιση στα λανθασμένα παραδείγματα σημαίνει ότι υπάρχει μεγάλη επικάλυψη στις λανθασμένες προβλέψεις των συστημάτων εξαγωγής του βασικού επιπέδου, “φρενάροντας” την αναμενόμενη πτώση της ακρίβειας σε μετα-επίπεδο, κατά το συνδυασμό είτε μέσω *MVotM* είτε μέσω *PVotM*. Μικρή συσχέτιση στα σωστά παραδείγματα σημαίνει μικρή επικάλυψη στις σωστές προβλέψεις των συστημάτων σε βασικό επίπεδο, οδηγώντας σε μεγαλύτερη ανάκληση σε μετα-επίπεδο, ταυτόχρονα με ελαφρά άνοδο της ακρίβειας, πάντα κατά το συνδυασμό μέσω *MVotM* ή *PVotM*.

Από την άλλη πλευρά, προσθέτοντας το σύστημα των HMMs στο συνδυασμό των άλλων δύο συστημάτων για τα μαθήματα της επιστήμης υπολογιστών, επίσης είτε μέσω *MVotM*, είτε μέσω *PVotM*, αν και αυξάνεται η ανάκληση οδηγείται σε σημαντική πτώση η ακρίβεια. Αυτό οφείλεται στη χαμηλότερη συσχέτιση των λανθασμένων παραδειγμάτων ανάμεσα στα HMMs και τα άλλα δύο συστήματα. Μικρή συσχέτιση στα λανθασμένα παραδείγματα σημαίνει μικρή επικάλυψη στις λανθασμένες προβλέψεις των συστημάτων σε βασικό επίπεδο, οδηγώντας σε μεγάλη πτώση την ακρίβεια σε μετα-επίπεδο κατά το συνδυασμό μέσω *MVotM* ή *PVotM*.

Όσον αφορά το *MVotF*, τα καλύτερα αποτελέσματα, με βάση τη μετρική $F1$, επιτυγχάνονται μόνο κατά το συνδυασμό και των τριών συστημάτων. Για ζευγάρια

συστημάτων, τυχαία επιλογή λαμβάνει χώρα είτε στην περίπτωση διφορούμενων πεδίων, κάτι που σπάνια γίνεται σύμφωνα με το Σχήμα 4.4, είτε όταν ένα μόνο σύστημα δεν προβλέπει κάποιο πεδίο (αγνοούμενη πρόβλεψη). Στην τελευταία περίπτωση, τυχαία επιλογή λαμβάνει χώρα ανάμεσα στην τιμή “false”, που κωδικοποιεί την αγνοούμενη πρόβλεψη, και στο πεδίο που προβλέπεται από το δεύτερο σύστημα. Κατά την ψηφοφορία και των τριών συστημάτων μέσω *MVotF*, σπάνια συμβαίνει τυχαία επιλογή ανάμεσα στο “false” και σε κάποιο προβλεπόμενο πεδίο. Για να συνέβαινε κάτι τέτοιο, θα έπρεπε δύο από τα συστήματα να προέβλεπαν διαφορετικά πεδία, κάτι όχι συχνό σύμφωνα με το Σχήμα 4.4, ενώ το τρίτο σύστημα να μην έκανε πρόβλεψη (άρα η τιμή “false” θα συμμετείχε στην ψηφοφορία).

Ο Πίνακας 4.9 δείχνει επίσης ότι ο συνδυασμός των HMMs με $(LP)^2$ στις περισσότερες περιπτώσεις οδηγεί σε καλύτερο $F1$ από τους άλλους συνδυασμούς ζευγαριών συστημάτων, το οποίο $F1$ είναι αρκετά κοντά στο αντίστοιχο του συνδυασμού και των τριών συστημάτων. Κάτι τέτοιο σημαίνει ότι η συνεισφορά του συστήματος του BWI δεν είναι ιδιαίτερα σημαντική και μπορεί μερικώς να δικαιολογηθεί από τον Πίνακα 3.11 όπου παρατηρείται μεγάλη συσχέτιση μεταξύ του BWI του καλύτερου συστήματος εξαγωγής πληροφορίας για κάθε θεματική περιοχή. Το τελικό συμπέρασμα είναι ότι και για ζευγάρια συστημάτων του βασικού επιπέδου, το *PVotF* είναι το καλύτερο από όσα σχήματα ψηφοφορίας αξιολογήθηκαν.

4.7 Συμπεράσματα

Η ανάλυση της εξόδου των συστημάτων εξαγωγής πληροφορίας του βασικού επιπέδου και στις πέντε θεματικές περιοχές, έδειξε ότι περιπτώσεις διφορούμενων προβλέψεων, δηλαδή διαφορετικών πεδίων που προβλέπονται από διαφορετικά συστήματα για ένα τμήμα κειμένου, συμβαίνουν σε πολύ μικρό ποσοστό. Κατά συνέπεια, το μοναδικό είδος διαφορετικότητας που καλούμαστε να εκμεταλλευτούμε κατά το συνδυασμό σε μετα-επίπεδο είναι οι περιπτώσεις όπου κάποια συστήματα προβλέπουν το ίδιο πεδίο, ενώ κάποια άλλα δεν κάνουν κάποια πρόβλεψη. Ερμηνεύοντας την απουσία πρόβλεψης ως άρνηση πρόβλεψης, ή παραβλέποντάς την, επηρεάζει σημαντικά την απόδοση της εξαγωγής κατά το συνδυασμό των συστημάτων μέσω τεχνικών *ψηφοφορίας*.

Η χρήση *πλειοψηφικής ψηφοφορίας* δεν αποδείχτηκε ιδιαίτερα αποτελεσματική κατά το συνδυασμό συστημάτων, σύμφωνα με τη μετρική $F1$. Ερμηνεύοντας όμως κάθε απουσία πρόβλεψης ενός συστήματος ως άρνηση πρόβλεψης, οδήγησε σε πιο *ακριβή* αποτελέσματα εξαγωγής σε σχέση με τα συστήματα του βασικού επιπέδου. Αντίθετα,

παραβλέποντας την απουσία πρόβλεψης, οδήγησε σε αποτελέσματα εξαγωγής με μεγαλύτερη *ανάκληση* σε μετα-επίπεδο. Από την άλλη πλευρά, θα ήταν πιο ενδιαφέρον να *μαθαίναμε* ποιο είναι το σωστό πεδίο για ένα τμήμα κειμένου, δοθέντων των πεδίων που προβλέπονται από τα συστήματα του βασικού επιπέδου.

Η χρήση *πιθανοτικής ψηφοφορίας* αποδείχτηκε αποτελεσματική στις περισσότερες θεματικές περιοχές, μόνο εφόσον τεθεί ένα όριο στην τελική πιθανότητα για την αποδοχή ή όχι μιας πρόβλεψης. Εάν δεν τεθεί ένα τέτοιο κατώφλι και παραβλέποντας την απουσία πρόβλεψης ενός συστήματος, η πιθανοτική ψηφοφορία συμπεριφέρεται παρόμοια με την αντίστοιχη περίπτωση της πλειοψηφικής ψηφοφορίας. Από την άλλη πλευρά, θα ήταν πιο ενδιαφέρον να *μαθαίναμε* ποιο είναι το σωστό πεδίο για ένα τμήμα κειμένου, δοθέντων των πεδίων και των αντίστοιχων πιθανοτήτων που προέρχονται από τα συστήματα του βασικού επιπέδου. Με άλλα λόγια, θα ήταν επιθυμητό να *μαθαίναμε* τις κατάλληλες συσχετίσεις μεταξύ των πιθανοτήτων στις προβλέψεις των συστημάτων του βασικού επιπέδου, προσδοκώντας σε καλύτερα αποτελέσματα στην εξαγωγή πληροφορίας σε μετα-επίπεδο.

Η σύγκριση όλων των τεχνικών ψηφοφορίας που ορίστηκαν σε αυτό το κεφάλαιο ανέδειξε ως νικήτρια στις τρεις από τις πέντε θεματικές περιοχές την πιθανοτική ψηφοφορία με χρήση κατωφλίου αποδοχής/απόρριψης προβλέψεων. Στις άλλες δύο περιοχές αναδείχτηκε νικήτρια η χρήση πλειοψηφικής ψηφοφορίας όπου κάθε απουσία πρόβλεψης ερμηνεύεται ως άρνηση πρόβλεψης. Συγκριτικό πλεονέκτημα, όμως, της πιθανοτικής ψηφοφορίας με χρήση κατωφλίου αποδοχής προβλέψεων, σε σχέση με τις άλλες τεχνικές, αποτελεί το γεγονός ότι δεν χειροτέρεψε σε μετα-επίπεδο τα αποτελέσματα εξαγωγής του καλύτερου συστήματος του βασικού επιπέδου, σε καμία από τις πέντε θεματικές περιοχές. Η πιθανοτική ψηφοφορία επίσης πέτυχε καλύτερα αποτελέσματα εξαγωγής από ότι η πολυστρατηγική μάθηση [48].

Τέλος, σε δύο θεματικές περιοχές (αγγελίες εργασίας και ανακοινώσεις σεμιναρίων) όπου παρατηρείται μεγάλος βαθμός συσχέτισης στις προβλέψεις των συστημάτων του βασικού επιπέδου, η χρήση ψηφοφορίας δεν οδήγησε σε ιδιαίτερα ικανοποιητικά αποτελέσματα στην εξαγωγή πληροφορίας σε μετα-επίπεδο, σε σχέση με τις υπόλοιπες τρεις θεματικές περιοχές. Με τη χρήση *μηχανικής μάθησης* σε μετα-επίπεδο αντί της χρήσης ψηφοφορίας, ευελπιστούμε στην επίτευξη καλύτερων αποτελεσμάτων εξαγωγής σε μετα-επίπεδο.

ΚΕΦΑΛΑΙΟ 5

ΣΥΝΔΥΑΣΜΟΣ ΣΥΣΤΗΜΑΤΩΝ ΕΞΑΓΩΓΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΜΕ ΧΡΗΣΗ ΣΥΣΣΩΡΕΥΜΕΝΗΣ ΓΕΝΙΚΕΥΣΗΣ

Αντίθετα με τη χρήση απλής ψηφοφορίας για το συνδυασμό συστημάτων εξαγωγής πληροφορίας, μια αρκετά πιο ενδιαφέρουσα ιδέα είναι η χρήση *μηχανικής μάθησης* σε μετα-επίπεδο. Εξετάζοντας ξανά το συσσωρευμένο σχεδιότυπο του Πίνακα 4.2, αναρωτιόμαστε για το εάν μπορούμε να *μάθουμε* να προβλέπουμε το σωστό πεδίο για κάθε τμήμα κειμένου, από τις προβλέψεις των συστημάτων που είναι διαθέσιμα σε βασικό επίπεδο. Ένα απλό κίνητρο για την προτίμηση μάθησης, αντί της ψηφοφορίας, για το συνδυασμό διαφορετικών συστημάτων, είναι ότι η ψηφοφορία αδυνατεί να χειριστεί περιπτώσεις λανθασμένων προβλέψεων από την πλειοψηφία των συστημάτων. Για παράδειγμα, εάν ένα σύστημα προβλέψει σωστά το πεδίο *ram* (χωρητικότητα κύριας μνήμης υπολογιστή) για ένα υποθετικό τμήμα κειμένου “1 GB” ενώ τα υπόλοιπα προβλέψουν λανθασμένα το πεδίο *HDcapacity* (χωρητικότητα σκληρού δίσκου), τότε επιλέγεται η τελευταία τιμή κατά την ψηφοφορία. Γι’ αυτό θα ήταν επιθυμητό με τη χρήση μάθησης σε μετα-επίπεδο να προκύψει ένας κανόνας ταξινόμησης της μορφής: *εάν το σύστημα* E^1 *προβλέψει το πεδίο “ram” ενώ τα υπόλοιπα συστήματα προβλέψουν “HDcapacity”, τότε το σωστό πεδίο είναι “ram”*.

Στην Ενότητα 5.1 προτείνεται η κατασκευή διανυσμάτων χαρακτηριστικών σε μετα-επίπεδο με βάση τις προβλέψεις των συστημάτων του βασικού επιπέδου, όπως περιέχονται στο συσσωρευμένο σχεδιότυπο. Στις Ενότητες 5.2 και 5.3 παρουσιάζεται μια νέα μέθοδος για το συνδυασμό συστημάτων εξαγωγής πληροφορίας η οποία βασίζεται στη χρήση *συσσωρευμένης γενίκευσης* και χρησιμοποιεί απλές ονομαστικές τιμές στις προβλέψεις των συστημάτων. Η Ενότητα 5.4 περιγράφει τα καινοτομικά χαρακτηριστικά της προτεινόμενης μεθοδολογίας σε σχέση με τη συσσωρευμένη γενίκευση για κοινά προβλήματα ταξινόμησης. Οι Ενότητες 5.5 και 5.6 παρουσιάζουν την εφαρμογή της νέας μεθοδολογίας με χρήση πιθανοτήτων ορθότητας στις προβλέψεις των συστημάτων. Στην Ενότητα 5.7 αξιολογείται εμπειρικά η προτεινόμενη μεθοδολογία συσσωρευμένης γενίκευσης, ενώ η Ενότητα 5.8 συνοψίζει τα συμπεράσματα του κεφαλαίου αυτού.

5.1 Κατασκευή διανυσμάτων χαρακτηριστικών σε μετα-επίπεδο

Στα πλαίσια της χρήσης μάθησης θα χρειαστεί να εκπαιδεύσουμε ένα κοινό ταξινομητή σε μετα-επίπεδο, παρέχοντας ως δεδομένα εκπαίδευσης ένα σύνολο *διανυσμάτων χαρακτηριστικών*. Η ιδέα που προτείνεται σε αυτή τη διατριβή είναι η δημιουργία ενός διανύσματος χαρακτηριστικών για κάθε γραμμή του συσσωρευμένου σχεδιοτύπου, δηλαδή για κάθε τμήμα κειμένου που έχει αναγνωριστεί από τουλάχιστον ένα σύστημα του βασικού επιπέδου. Τα χαρακτηριστικά στα νέα διανύσματα θα είναι οι προβλέψεις των συστημάτων του βασικού επιπέδου. Ο Πίνακας 5.1 δείχνει τα νέα διανύσματα που δημιουργούνται από το συσσωρευμένο σχεδιότυπο του Πίνακα 4.2.

Πίνακας 5.1 Διανύσματα χαρακτηριστικών σε μετα-επίπεδο που έχουν δημιουργηθεί από το συσσωρευμένο σχεδιότυπο του Πίνακα 4.2.

s, e	$t(s,e)$	*Διανύσματα χαρακτηριστικών χρησιμοποιώντας ονομαστικές τιμές*		
		Έξοδος από E^1	Έξοδος από E^2	Κλάση
47, 49	TransPort ZX	model,	manuf,	model
56, 58	15"	screenSize,	?,	screenSize
59, 60	TFT	screenType,	screenType,	screenType
63, 66	Intel<b>Pentium	?,	procName,	false
63, 67	Intel<b>Pentium III	procName,	?,	procName
67, 69	600 MHz	procSpeed,	procSpeed,	procSpeed
76, 78	256 MB	ram,	ram,	ram
81, 83	1 GB	ram,	HDcapacity,	false
86, 88	40 GB	?,	HDcapacity,	HDcapacity

Κάθε απουσία πρόβλεψης για ένα τμήμα κειμένου αντιπροσωπεύεται με “?” στα διανύσματα χαρακτηριστικών. Εάν ένα τμήμα κειμένου δεν υπάρχει στο χειρονακτικά συμπληρωμένο σχεδιότυπο, τότε το χαρακτηριστικό κλάσης (τελευταία στήλη στον Πίνακα 5.1) του αντίστοιχου διανύσματος παίρνει την τιμή “false”, υποδεικνύοντας ότι κανένα πεδίο δεν πρέπει να προβλεφθεί για το τμήμα αυτό. Κατά τη διαδικασία επαλήθευσης (runtime), η τιμή του χαρακτηριστικού κλάσης που πρέπει να προβλεφθεί από τον ταξινομητή.

Αυτό που απομένει είναι η κατασκευή του πλήρους συνόλου των διανυσμάτων χαρακτηριστικών για την εκπαίδευση του ταξινομητή σε μετα-επίπεδο. Αυτό γίνεται με χρήση *διασταυρωμένης επικύρωσης* στο σύνολο των δεδομένων εκπαίδευσης, όπως περιγράφεται στην Ενότητα 2.3. Στην εξαγωγή πληροφορίας, όμως, χειριζόμαστε συλλογές κειμένων, επισημειωμένων με παραδείγματα σχετικών πεδίων και όχι διανύσματα χαρακτηριστικών όπως στην κοινή ταξινόμηση. Επομένως, η διαδικασία της διασταυρωμένης επικύρωσης πρέπει να πραγματοποιηθεί στα επισημειωμένα

κείμενα εκπαίδευσης και όχι σε διανύσματα χαρακτηριστικών όπως συμβαίνει στη συσσωρευμένη γενίκευση για προβλήματα ταξινόμησης. Αυτή η ανομοιογένεια ανάμεσα στα δεδομένα του βασικού επιπέδου, που αποτελούνται από συλλογές κειμένων, και στα δεδομένα του μετα-επιπέδου, που αποτελούνται από διανύσματα χαρακτηριστικών, αποτελεί το κύριο χαρακτηριστικό στη χρήση συσσωρευμένης γενίκευσης για το συνδυασμό διαφορετικών συστημάτων εξαγωγής πληροφορίας.

5.2 Συσσωρευμένη γενίκευση με χρήση ονομαστικών τιμών

Η βασική ιδέα πίσω από τον εφαρμογή συσσωρευμένης γενίκευσης στο πρόβλημα της εξαγωγής πληροφορίας, είναι η εκπαίδευση ενός ταξινομητή σε μετα-επίπεδο, βάσει της εξόδου ενός συνόλου συστημάτων εξαγωγής πληροφορίας που είναι διαθέσιμα σε βασικό επίπεδο. Διασταυρωμένη επικύρωση πραγματοποιείται στα επισημειωμένα κείμενα εκπαίδευσης για την δημιουργία του συνόλου διανυσμάτων χαρακτηριστικών που θα χρησιμοποιηθούν για την εκπαίδευση του ταξινομητή.

Στο *j*-βήμα, $j = 1..J$ της διαδικασίας διασταυρωμένης επικύρωσης, οι N αλγόριθμοι μηχανικής μάθησης $L^1...L^N$ εφαρμόζονται στο υποσύνολο $D \setminus D^j$ των κειμένων εκπαίδευσης και τα εκπαιδευμένα συστήματα εξαγωγής πληροφορίας $E^1(j)...E^N(j)$ εφαρμόζονται στο σύνολο D^j των κειμένων επαλήθευσης. Για κάθε κείμενο d στο σύνολο D^j, έστω $T^1...T^N$ τα συσσωρευμένα σχεδιότυπα που έχουν συμπληρωθεί για το d από τα συστήματα $E^1(j)...E^N(j)$ αντίστοιχα. Ένα συσσωρευμένο σχεδιότυπο **ΣΣ** δημιουργείται από τα $T^1...T^N$. Ένα καινούριο διάνυσμα χαρακτηριστικών κατασκευάζεται για κάθε καταχωρημένη γραμμή του συσσωρευμένου σχεδιοτύπου, όπως φαίνεται στον Πίνακα 5.1.

Κάθε νέο διάνυσμα προστίθεται στο σύνολο MD^j των δεδομένων του μετα-επιπέδου. Στο τέλος ολόκληρης της διαδικασίας διασταυρωμένης επικύρωσης, η ένωση $MD = \cup MD^j$ αποτελεί το σύνολο των δεδομένων του μετα-επιπέδου, δηλαδή το σύνολο των διανυσμάτων χαρακτηριστικών του μετα-επιπέδου, το οποίο χρησιμοποιείται από έναν αλγόριθμο μηχανικής μάθησης L^M για την εκπαίδευση του ταξινομητή C^M. Τελικά, οι N αλγόριθμοι μηχανικής μάθησης $L^1...L^N$ εφαρμόζονται ξανά σε ολόκληρο το σύνολο D των κειμένων εκπαίδευσης, για την τελική εκπαίδευση των συστημάτων $E^1...E^N$ του βασικού επιπέδου που θα χρησιμοποιηθούν κατά τη διαδικασία επαλήθευσης. Το Σχήμα 5.1 δείχνει μια αλγοριθμική περιγραφή της προτεινόμενης μεθόδου συσσωρευμένης γενίκευσης για εξαγωγή πληροφορίας.

procedure Συσσωρευμένη Γενίκευση για Εξαγωγή Πληροφορίας (D , J , $L^1 ... L^N$, L^M)
begin
  $D^1 ... D^J$ = διαχωρισμός του D σε J συλλογές κειμένου σχεδόν ίδιου μεγέθους
  for j = 1 **to** J **do begin**
    MD^j = {}
    for i = 1 **to** N **do**
      $E^i(j)$ = το σύστημα εξαγωγής πληροφορίας που προκύπτει από την εφαρμογή του αλγορίθμου L^i στο σύνολο κειμένων $D \setminus D^j$
      foreach κείμενο d στο σύνολο D^j **do**
        begin
        for i = 1 **to** N **do**
          T^i = το σχεδιότυπο, συμπληρωμένο από το σύστημα $E^i(j)$
        ΣΣ = *Δημιουργία του συσσωρευμένου σχεδιοτύπου από* (d , $T^1 ... T^N$)
        foreach καταχώρηση, δηλ., για κάθε τμήμα κειμένου $t(s,e)$, στο *ΣΣ* **do**
          begin
            for i = 1 **to** N **do**
              $f^i \in \{f^1, ... f^Q, "?"\}$ = το πεδίο από το σύστημα $E^i(j)$ για το $t(s,e)$
            $f \in \{f^1, ... f^Q, false\}$ = το σωστό πεδίο για το τμήμα κειμένου $t(s,e)$
            $MD^j = MD^j \cup$ διάνυσμα $< f^1, ... f^N, f >$
          end
        end
  end // τέλος της διαδικασίας διασταυρωμένης επικύρωσης
  $MD = \cup MD^j$, $j = 1 ... J$
  C^M = ταξινομητής μετα-επιπέδου που προκύπτει από την εφαρμογή του αλγορίθμου L^M στο σύνολο MD των διανυσμάτων χαρακτηριστικών
  // *Εκπαίδευση των συστημάτων εξαγωγής πληροφορίας σε βασικό επίπεδο*
  for i = 1 **to** N **do**
    E^i = το σύστημα εξαγωγής πληροφορίας βασικού επιπέδου που προκύπτει από την εφαρμογή του αλγορίθμου L^i στο σύνολο D των κειμένων εκπαίδευσης
end

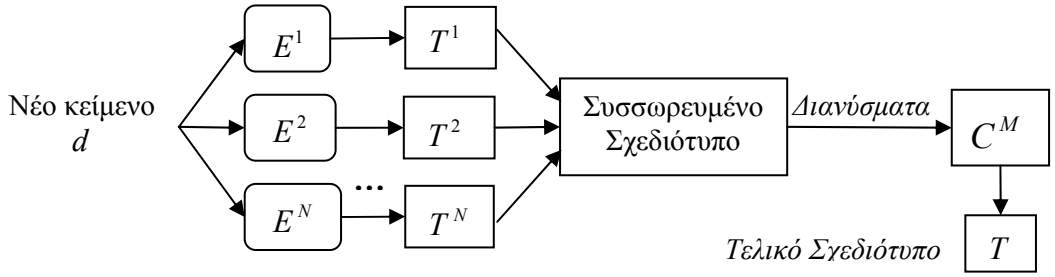
Σχήμα 5.1 Η προτεινόμενη μεθοδολογία συσσωρευμένης γενίκευσης για το συνδυασμό συστημάτων εξαγωγής πληροφορίας.

Τα διανύσματα χαρακτηριστικών στο νέο σύνολο MD των δεδομένων του μετα-επιπέδου ανήκουν σε $Q+1$ κλάσεις, όπου Q είναι ο αριθμός των σχετικών πεδίων σε μια θεματική περιοχή συν την τιμή “false”. Ένα διάνυσμα που ταξινομείται ως “false” υποδεικνύει ότι το αντίστοιχο τμήμα κειμένου $t(s,e)$ δεν υπάρχει στο χειρονακτικά συμπληρωμένο σχεδιότυπο για το κείμενο και για το λόγο αυτό κανένα από τα συστήματα του βασικού επιπέδου δεν θα έπρεπε να το είχε αναγνωρίσει ως σχετικό, προβλέποντας κάποιο πεδίο γι’ αυτό (όπως για παράδειγμα το τμήμα κειμένου “1 GB” στον Πίνακα 5.1). Επίσης, σε αντιστοιχία με την πλειοψηφική ψηφοφορία, οι αγνοούμενες τιμές (“?” στον Πίνακα 5.1) μπορούν να αντικατασταθούν με τιμές “false”, υποδεικνύοντας άρνηση πρόβλεψης. Στην περίπτωση αυτή, όλα τα χαρακτηριστικά των

διανυσμάτων σε μετα-επίπεδο, συμπεριλαμβανομένου και του τελευταίου χαρακτηριστικού κλάσης, θα παίρνουν τιμές από το ίδιο σύνολο τιμών $\{f^1...f^Q, false\}$.

5.3 Χρήση συσσωρευμένης γενίκευσης κατά την επαλήθευση

Το Σχήμα 5.2 δείχνει τη λειτουργία κατά τη διαδικασία επαλήθευσης της προτεινόμενης μεθοδολογίας συσσωρευμένης γενίκευσης για εξαγωγή πληροφορίας.



Σχήμα 5.2 Η προτεινόμενη μεθοδολογία συσσωρευμένης γενίκευσης για εξαγωγή πληροφορίας κατά τη διαδικασία επαλήθευσης

Δοθέντος ενός νέου κειμένου d, τα εκπαιδευμένα συστήματα $E^1...E^N$ του βασικού επιπέδου χρησιμοποιούνται για να βρουν παραδείγματα σχετικών πεδίων στο d και να συμπληρώσουν τα αντίστοιχα σχεδιότυπα $T^1...T^N$. Ένα συσσωρευμένο σχεδιότυπο δημιουργείται στη συνέχεια από τα $T^1...T^N$. Για κάθε γραμμή στο συσσωρευμένο σχεδιότυπο, δηλαδή για κάθε διακεκριμένο τμήμα κειμένου $t(s,e)$, δημιουργείται ένα διάνυσμα χαρακτηριστικών από τα πεδία που προβλέπουν τα συστήματα $E^1...E^N$ για το $t(s,e)$ (απουσία πρόβλεψης από κάποιο σύστημα μεταφράζεται ως "?" ή με "false"). Τα νέα διανύσματα ταξινομούνται τελικά από τον ταξινομητή C^M σε μια από $Q+1$ προκαθορισμένες κατηγορίες $\{f^1...f^Q, false\}$. Εάν ένα διάνυσμα ταξινομείται σε ένα από τα σχετικά πεδία $\{f^1...f^Q\}$, τότε το αντίστοιχο παράδειγμα $< t(s,e), f >$ εισάγεται στο τελικό σχεδιότυπο για το κείμενο d. Αλλιώς ("false" πρόβλεψη) το αντίστοιχο παράδειγμα αγνοείται και δεν εισάγεται στο τελικό σχεδιότυπο. Για παράδειγμα, εάν υποθέσουμε ότι η τελευταία στήλη στον Πίνακα 5.1 έχει συμπληρωθεί από τον ταξινομητή C^M, τότε για τα δύο διανύσματα που έχουν ταξινομηθεί ως "false", τα αντίστοιχα παραδείγματα δεν θα εισαχθούν στο τελικό σχεδιότυπο για τη σελίδα.

5.4 Ιδιαιτερότητες της προτεινόμενης μεθοδολογίας

Η βασική διαφορά ανάμεσα στη συσσωρευμένη γενίκευση για εξαγωγή πληροφορίας και στην συσσωρευμένη γενίκευση για κοινά προβλήματα ταξινόμησης, είναι ότι η

διαδικασία διασταυρωμένης επικύρωσης στην πρώτη περίπτωση εφαρμόζεται σε αρχεία κειμένου, μαζί με τα αντίστοιχα χειρονακτικά συμπληρωμένα σχεδιότυπα, αντί σε διανύσματα χαρακτηριστικών όπως στην δεύτερη περίπτωση. Αυτό αναιρεί τον περιορισμό της πραγματοποίησης κοινής ταξινόμησης σε βασικό επίπεδο. Αναιρείται δηλαδή ο περιορισμός ότι τα δεδομένα του βασικού επιπέδου πρέπει να είναι κι αυτά διανύσματα χαρακτηριστικών, επιτρέποντας την εφαρμογή της συσσωρευμένης γενίκευσης και στο πρόβλημα της εξαγωγής πληροφορίας. Βέβαια, ως αποτέλεσμα της αντικατάστασης των διανυσμάτων χαρακτηριστικών του βασικού επιπέδου από επισημειωμένα κείμενα, προκύπτουν μια σειρά από ενδιαφέροντα ζητήματα τα οποία χρήζουν αναφοράς και περεταίρω διερεύνησης.

Οι αλγόριθμοι $L^1 ... L^N$ που χρησιμοποιούνται στο βασικό επίπεδο είναι σχεδιασμένοι για εξαγωγή πληροφορίας, ενώ ο αλγόριθμος L^M που χρησιμοποιείται σε μετα-επίπεδο είναι σχεδιασμένος για κοινά προβλήματα ταξινόμησης και δεν μπορεί να είναι ένας από τους αλγορίθμους $L^1 ... L^N$. Αντίθετα, στη συσσωρευμένη γενίκευση για προβλήματα ταξινόμησης, οι αλγόριθμοι σε βασικό και μετα-επίπεδο είναι σχεδιασμένοι για ταξινόμηση και επομένως ο L^M μπορεί να είναι κάποιος από τους $L^1 ... L^N$.

Το μέγεθος του συνόλου των δεδομένων σε μετα-επίπεδο, δηλαδή ο αριθμός των διανυσμάτων, δεν είναι εκ των προτέρων γνωστός στη συσσωρευμένη γενίκευση για εξαγωγή πληροφορίας, σε αντίθεση με τη συσσωρευμένη γενίκευση για ταξινόμηση όπου υπάρχει ένα προς ένα αντιστοιχία ανάμεσα στα διανύσματα χαρακτηριστικών του βασικού και του μετα-επιπέδου. Στην εξαγωγή πληροφορίας, το μέγεθος του συνόλου δεδομένων σε μετα-επίπεδο προσδιορίζεται από τις προβλέψεις των συστημάτων βασικού επιπέδου, δηλαδή από τα σχεδιότυπα $T^1 ... T^N$ που σχηματίζουν στη συνέχεια το συσσωρευμένο σχεδιότυπο, όπως φαίνεται και στο Σχήμα 5.1.

Το γεγονός της ένα προς ένα αντιστοιχίας των διανυσμάτων χαρακτηριστικών του βασικού και του μετα-επιπέδου στη συσσωρευμένη γενίκευση για την κοινή ταξινόμηση, σημαίνει αυτόματα ότι οι τιμές του χαρακτηριστικού κλάσης παραμένουν οι ίδιες σε βασικό και σε μετα-επίπεδο. Κάτι τέτοιο δεν ισχύει κατά κανόνα στη συσσωρευμένη γενίκευση για εξαγωγή πληροφορίας, όπου ένα επισημειωμένο παράδειγμα $< t(s,e), f >$ σε ένα κείμενο εκπαίδευσης ενδέχεται να μην αντιστοιχεί σε διάνυσμα χαρακτηριστικών σε μετα-επίπεδο. Αυτό συμβαίνει όταν κανένα από τα διαθέσιμα συστήματα του βασικού επιπέδου δεν προβλέψει κάποιο πεδίο για το τμήμα $t(s,e)$, οπότε δεν θα δημιουργηθεί διάνυσμα χαρακτηριστικών για το τμήμα αυτό στο σύνολο διανυσμάτων του μετα-επιπέδου. Το αποτέλεσμα θα είναι η απώλεια, σε μετα-επίπεδο, μερικών

επισημειωμένων παραδειγμάτων σχετικών πεδίων του βασικού επιπέδου. Παρά την απώλεια αυτή, αισιοδοξούμε για την επιτυχία του ταξινομητή που θα εκπαιδευτεί σε μετα-επίπεδο όσον αφορά τη βελτίωση των αποτελεσμάτων εξαγωγής των συστημάτων του βασικού επιπέδου. Επίσης, η παραπάνω παρατήρηση υποδεικνύει ότι θα πρέπει να προτιμώνται συστήματα με μεγαλύτερη *ανάκληση* κατά το συνδυασμό μέσω συσσωρευμένης γενίκευσης, ευελπιστώντας στην ελάττωση της απώλειας πληροφορίας σε μετα-επίπεδο.

Επιπλέον, στη συσσωρευμένη γενίκευση για ταξινόμηση, ένα διάνυσμα χαρακτηριστικών σε μετα-επίπεδο δεν περιέχει αγνοούμενες τιμές, αφού κάθε ταξινομητής του βασικού επιπέδου πάντα προβλέπει μια ονομαστική τιμή κλάσης ή μια πιθανοτική κατανομή για όλες τις κλάσεις. Αντιθέτως, ένα διάνυσμα χαρακτηριστικών μετα-επιπέδου στη συσσωρευμένη γενίκευση για εξαγωγή πληροφορίας μπορεί να περιλαμβάνει αγνοούμενες τιμές, όπως φαίνεται και στον Πίνακα 5.1, αφού κάποιο σύστημα του βασικού επιπέδου μπορεί να μην έχει προβλέψει κάποιο πεδίο για ένα τμήμα κειμένου που έχει αναγνωριστεί από άλλο σύστημα ή συστήματα.

Ένα ζήτημα στη συσσωρευμένη γενίκευση για εξαγωγή πληροφορίας είναι η επιλογή της τιμής J κατά τη διασταυρωμένη επικύρωση, όπως φαίνεται και στο Σχήμα 5.1. Η επιλογή του J εξαρτάται συνήθως από το μέγεθος των δεδομένων εκπαίδευσης, καθώς και από το κόστος εκπαίδευσης. Η επιλογή του J επηρεάζει επίσης και τη διαφορά στον αριθμό των κειμένων που χρησιμοποιούνται για την εκπαίδευση των συστημάτων του βασικού επιπέδου και του αριθμού των κειμένων εκείνων από τα οποία κατασκευάζονται τα διανύσματα χαρακτηριστικών για την εκπαίδευση των ταξινομητών σε μετα-επίπεδο. Σύμφωνα με το Σχήμα 5.1, τα συστήματα $E^1...E^N$ σε βασικό επίπεδο που πρόκειται να χρησιμοποιηθούν κατά την επαλήθευση, επανεκπαιδεύονται σε ολόκληρη τη συλλογή D των επισημειωμένων κειμένων εκπαίδευσης. Από την άλλη πλευρά, τα διανύσματα χαρακτηριστικών σε κάθε *j*-βήμα της διασταυρωμένης επικύρωσης, δημιουργούνται από τις προβλέψεις των συστημάτων $E^1(j)...E^N(j)$, όπως αυτά εκπαιδεύονται σε ένα μικρότερο τμήμα D/D^j των κειμένων. Όσο μεγαλύτερη είναι η τιμή για το J, τόσο μικρότερο το μέγεθος του D^j και κατά συνέπεια μικρότερη και η διαφορά ανάμεσα στα κείμενα εκπαίδευσης που χρησιμοποιούνται για την εκπαίδευση των συστημάτων $E^1(j)...E^N(j)$ και του πλήρους συνόλου D που χρησιμοποιείται για την εκπαίδευση των συστημάτων $E^1...E^N$.

Για να γίνουν περισσότερο κατανοητά τα παραπάνω, για μια συλλογή 40 κειμένων εκπαίδευσης, τα τελικά συστήματα εξαγωγής πληροφορίας $E^1...E^N$ εκπαιδεύονται σε

ολόκληρη τη συλλογή αυτή. Υποθέτοντας μια διαδικασία διασταυρωμένης επικύρωσης 5 βημάτων, τότε σε κάθε *j*-βήμα, τα συστήματα $E^1(j)...E^N(j)$ εκπαιδεύονται σε 32 κείμενα. Εναλλακτικά, η χρήση μεγαλύτερης τιμής για το J θα είχε ως αναπόφευκτη συνέπεια μεγαλύτερο υπολογιστικό κόστος κατά την εκπαίδευση των συστημάτων. Βέβαια, η επιλογή μεγαλύτερης τιμής για το J δε σημαίνει απαραίτητα και βελτίωση της απόδοσης σε μετα-επίπεδο. Ο Breiman [14] έχει δείξει ότι για αρκετά προβλήματα συσσωρευμένης παλινδρόμησης, η χρήση συσσωρευμένης γενίκευσης 10 βημάτων δουλεύει καλύτερα από τη χρήση V βημάτων, όπου V ο αριθμός των διανυσμάτων εκπαίδευσης από τα οποία τα $V-1$ χρησιμοποιούνται για την εκπαίδευση σε κάθε βήμα και το εναπομένον διάνυσμα για επαλήθευση.

Μια διαφορά, επίσης, ανάμεσα στη συσσωρευμένη γενίκευση για εξαγωγή πληροφορίας και στην αντίστοιχη για προβλήματα ταξινόμησης, αφορά το θέμα της *διαστρωμάτωσης (stratification)* κατά τη διασταυρωμένη επικύρωση. Στην κοινή ταξινόμηση, όπου η διασταυρωμένη επικύρωση λαμβάνει χώρα σε διανύσματα χαρακτηριστικών, διατηρείται μια παρόμοια κατανομή των σχετικών κλάσεων σε κάθε βήμα της διαδικασίας. Στην εξαγωγή πληροφορίας, από την άλλη, τυπικά υπάρχει μια διαφορετική κατανομή των παραδειγμάτων πεδίων σε κάθε κείμενο και γι' αυτό είναι αρκετά δύσκολο να προσεγγιστεί μια παρόμοια κατανομή των παραδειγμάτων σχετικών πεδίων σε κάθε βήμα. Μια εναλλακτική προσέγγιση θα ήταν να θεωρήσουμε κάθε πεδίο ξεχωριστά κατά το συνδυασμό, όπως προτείνεται στην εργασία [48]. Κάτι τέτοιο θα σήμαινε τη διενέργεια ξεχωριστών διαδικασιών διασταυρωμένης επικύρωσης, μια για κάθε σχετικό πεδίο, όπου θα υπήρχε όμως η προφανής δυνατότητα να προσεγγιστεί καλύτερα ο αριθμός των παραδειγμάτων του πεδίου σε κάθε βήμα της διαδικασίας. Το αρνητικό αντιστάθμισμα για μια τέτοια επιλογή θα ήταν να αγνοηθούν οι περιπτώσεις των διφορούμενων πεδίων που προβλέπονται από διαφορετικά συστήματα για ένα τμήμα κειμένου (όπως το τμήμα "TransPort ZX" στον Πίνακα 5.1).

5.5 Συσσωρευμένη γενίκευση με χρήση πιθανοτήτων

Μια άμεση επέκταση της προτεινόμενης μεθοδολογίας συσσωρευμένης γενίκευσης για εξαγωγή πληροφορίας είναι η χρήση αριθμητικών *τιμών εμπιστοσύνης (confidence values)* στην έξοδο των συστημάτων του βασικού επιπέδου, οι οποίες έχουν μετατραπεί σε *πιθανότητες ορθότητας (probabilities of correctness)*. Η νέα μεθοδολογία είναι η εξής:

- Αντί για την πρόβλεψη ενός από τα Q πεδία για ένα τμήμα κειμένου $t(s,e)$, κάθε σύστημα παράγει μια αριθμητική τιμή εμπιστοσύνης c^k για το προβλεπόμενο πεδίο

f^k. Αυτό μοντελοποιείται από ένα διάνυσμα Q αριθμητικών χαρακτηριστικών από μηδενικές τιμές, εκτός από την *k*-θέση όπου το c^k εμφανίζεται, δηλαδή $<0,...,c^k,...0>$. Εάν ένα σύστημα δεν προβλέψει κάποιο πεδίο για ένα τμήμα $t(s,e)$, τότε όλα τα Q χαρακτηριστικά αποτελούνται από μηδενικές τιμές.

- Κάθε διάνυσμα Q χαρακτηριστικών μετατρέπεται σε ένα νέο διάνυσμα $<0,...p^k,...0>$, όπου η τιμή p^k αντιστοιχεί στην τιμή c^k και εκφράζει την πιθανότητα ορθότητας της συγκεκριμένης πρόβλεψης. Η διαδικασία μετατροπής βαθμών εμπιστοσύνης στις προβλέψεις ενός συστήματος σε πιθανότητας ορθότητας περιγράφεται με περισσότερη λεπτομέρεια στην εργασία [48].
- Τελικά, τα διανύσματα εξόδου των συστημάτων $E^1...E^N$ για ένα τμήμα κειμένου $t(s,e)$, ενώνονται σε ένα διάνυσμα $N*Q$ χαρακτηριστικών, το οποίο συμπληρώνεται από το σωστό πεδίο για το $t(s,e)$ με βάση το χειρονακτικά συμπληρωμένο σχεδιότυπο. Εάν δεν υπάρχει καταχώρηση για το $t(s,e)$ στο χειρονακτικά συμπληρωμένο σχεδιότυπο, τότε το σωστό πεδίο για το $t(s,e)$ που συμπληρώνει το νέο διάνυσμα των $N*Q$ χαρακτηριστικών είναι η τιμή “false”.

Ο Πίνακας 5.2 δείχνει ένα επεξηγηματικό παράδειγμα των νέων διανυσμάτων χαρακτηριστικών σε μετα-επίπεδο, χρησιμοποιώντας πιθανοτικές τιμές ορθότητας.

Πίνακας 5.2 Διανύσματα χαρακτηριστικών σε μετα-επίπεδο που έχουν δημιουργηθεί από το συσσωρευμένο σχεδιότυπο του Πίνακα 4.2 και βασίζονται στην χρήση πιθανοτικών εκτιμήσεων ορθότητας.

s,e	$t(s,e)$	*Διανύσματα χαρακτηριστικών, χρησιμοποιώντας πιθανοτικές εκτιμήσεις*		
		Έξοδος από E^1	Έξοδος από E^2	Κλάση
47, 49	TransPort ZX	0, 0, 0.92, 0, 0, 0, 0, 0,	0, 0.34, 0, 0, 0, 0, 0, 0,	model
56, 58	15"	0, 0, 0, 0, 0, 0, 0.83, 0,	0, 0, 0, 0, 0, 0, 0, 0,	screenSize
59, 60	TFT	0, 0, 0, 0, 0, 0, 0, 0.85,	0, 0, 0, 0, 0, 0, 0, 0.91,	screenType
63, 66	Intel<b>Pentium	0, 0, 0, 0, 0, 0, 0, 0,	0, 0, 0, 0.61, 0, 0, 0, 0,	false
63, 67	Intel<b>Pentium III	0, 0, 0, 0.67, 0, 0, 0, 0,	0, 0, 0, 0, 0, 0, 0, 0,	procName
67, 69	600 MHz	0, 0, 0, 0, 0.82, 0, 0, 0,	0, 0, 0, 0, 0.79, 0, 0, 0,	procSpeed
76, 78	256 MB	0, 0, 0, 0, 0, 0.91, 0, 0,	0, 0, 0, 0, 0, 0.77, 0, 0,	ram
81, 83	1 GB	0, 0, 0, 0, 0, 0.55, 0, 0,	0.89, 0, 0, 0, 0, 0, 0, 0,	false
86, 88	40 GB	0, 0, 0, 0, 0, 0, 0, 0,	0.65, 0, 0, 0, 0, 0, 0, 0,	HDcapacity

Η ίδια αναπαράσταση στα διανύσματα του Πίνακα 5.2, έχει επίσης χρησιμοποιηθεί στην επέκταση της συσσωρευμένης γενίκευσης για προβλήματα ταξινόμησης, όπως προτάθηκε στην εργασία [107]. Στην εργασία αυτή, όμως, η έξοδος κάθε ταξινομητή είναι μια πιθανοτική κατανομή σε όλες τις κλάσεις (πεδία). Τέτοιες κατανομές δεν παράγονται τυπικά από τα συστήματα εξαγωγής πληροφορίας. Δοθέντος ενός

τμήματος κειμένου $t(s,e)$, είτε ένα πεδίο f προβλέπεται για το $t(s,e)$ ή δεν προβλέπεται κανένα πεδίο. Επομένως, εκτός από τα χαρακτηριστικά του διανύσματος που αντιστοιχούν στα πεδία που έχουν προβλεφθεί, όλα τα άλλα χαρακτηριστικά παίρνουν μηδενικές τιμές, όπως φαίνεται και στον Πίνακα 5.2.

Στη συσσωρευμένη γενίκευση με χρήση πιθανοτήτων, μια αγνοούμενη τιμή από ένα σύστημα για ένα τμήμα κειμένου αναπαρίσταται από ένα διάνυσμα του οποίου όλα τα χαρακτηριστικά παίρνουν μηδενικές τιμές. Οι αγνοούμενες τιμές μπορούν να αντιμετωπιστούν με την επέκταση της εξόδου κάθε συστήματος με ένα επιπλέον χαρακτηριστικό το οποίο υποδεικνύει την πιθανότητα για την τιμή “false”. Ο συνολικός αριθμός των χαρακτηριστικών σε μετα-επίπεδο θα είναι τώρα $N(Q+1)$. Η τιμή του επιπλέον χαρακτηριστικού θα είναι συμπληρωματική της μη μηδενικής τιμής του διανύσματος των Q χαρακτηριστικών για κάθε σύστημα.

Για παράδειγμα, στην πρώτη γραμμή του Πίνακα 5.2, η τιμή για το επιπλέον χαρακτηριστικό του συστήματος E^1 θα είναι 0.08, ενώ κάθε αγνοούμενη τιμή συνεπάγεται τιμή 1 για το επιπλέον χαρακτηριστικό. Η σπουδαιότητα της αντιμετώπισης αγνοούμενων τιμών στη συσσωρευμένη γενίκευση για εξαγωγή πληροφορίας, είτε με χρήση ονομαστικών τιμών, είτε με χρήση πιθανοτήτων, μπορεί να αξιολογηθεί εμπειρικά με τη σύγκριση της απόδοσης των εκπαιδευμένων ταξινομητών που χρησιμοποιούν τα νέα διανύσματα χαρακτηριστικών, σε αντιπαράθεση με τους ίδιους ταξινομητές, όταν αυτοί έχουν εκπαιδευτεί με τα παλαιά διανύσματα που δεν χειρίζονται τις αγνοούμενες τιμές.

5.6 Μετατροπή βαθμού εμπιστοσύνης σε πιθανότητες ορθότητας στις προβλέψεις των συστημάτων εξαγωγής πληροφορίας

Οι πιθανότητες που χρησιμοποιούνται από τις τεχνικές ψηφοφορίας και συσσωρευμένης γενίκευσης δεν προέρχονται απ’ ευθείας από την έξοδο των συστημάτων του βασικού επιπέδου. Αλγόριθμοι οι οποίοι μαθαίνουν κανόνες εξαγωγής πληροφορίας για την αναγνώριση σχετικών τμημάτων μέσα σε κείμενα, προσδιορίζουν και μια μετρική για το βαθμό εμπιστοσύνης της ακρίβειας των κανόνων που μαθαίνουν. Ο βαθμός αυτός εμπιστοσύνης κάθε κανόνα εκχωρείται στη συνέχεια στα τμήματα κειμένου τα οποία ο κανόνας αναγνωρίζει ως σχετικά για ένα πεδίο μέσα στη σελίδα. Για παράδειγμα, στον αλγόριθμο $(LP)^2$ ο βαθμός εμπιστοσύνης μετριέται για κάθε κανόνα μέσω του αριθμού των λανθασμένων αναγνωρίσεων που γίνονται από τον κανόνα κατά την εκπαίδευση, ενώ ο RAPIER [16] χρησιμοποιεί μια μετρική η οποία ενσωματώνει πληροφορία για το μέγεθος ενός κανόνα και του αριθμού των

λανθασμένων αναγνωρίσεών του σε μια μονή εξίσωση. Για τα HMMs, ο αλγόριθμος *Viterbi* [115] μπορεί να χρησιμοποιηθεί για την εκχώρηση βαθμών εμπιστοσύνης στα τμήματα κειμένου που αναγνωρίζονται ως σχετικά.

Το κίνητρο για τη μετατροπή βαθμών εμπιστοσύνης σε πιθανότητες ορθότητας, είναι ότι δεν υπάρχει πάντα ένας ισχυρός δεσμός μεταξύ τους, αφού, όπως έχει ήδη αναφερθεί και στην παράγραφο 2.5.1, σε τμήματα κειμένου τα οποία έχουν λανθασμένα αναγνωριστεί ως σχετικά είναι δυνατόν να εκχωρηθούν μεγαλύτερες τιμές εμπιστοσύνης από άλλα τμήματα κειμένου που έχουν αναγνωριστεί ως σχετικά. Οι βαθμοί εμπιστοσύνης που παράγονται από διαφορετικά συστήματα μπορεί επίσης να μην είναι άμεσα συγκρίσιμοι μεταξύ τους. Για παράδειγμα, οι βαθμοί εμπιστοσύνης που παράγει ο αλγόριθμος BWI, είναι θετικοί αριθμοί, ενώ οι αντίστοιχοι των HMMs είναι αρνητικοί, εξαιτίας του αλγορίθμου *Viterbi* και των λογαρίθμων που χρησιμοποιεί. Εάν πραγματοποιηθεί ψηφοφορία μεταξύ των συστημάτων που εκπαιδεύονται με χρήση των προαναφερθέντων αλγορίθμων, τότε τα αποτελέσματα δεν θα είναι αξιόπιστα. Μια εναλλακτική προσέγγιση θα ήταν να γίνει κανονικοποίηση των βαθμών εμπιστοσύνης κάθε συστήματος στο διάστημα μεταξύ μηδέν και ένα, αλλά και πάλι δεν είναι καθόλου σίγουρο ότι οι βαθμοί εμπιστοσύνης (έστω κανονικοποιημένες) αντιστοιχούν σε πιθανότητες ορθότητας.

Επομένως, το ζητούμενο είναι να βρεθεί μεν ένας τρόπος κανονικοποίησης των βαθμών εμπιστοσύνης που παράγει κάθε σύστημα στο διάστημα μεταξύ μηδέν και ένα, αλλά από την άλλη οι κανονικοποιημένες αυτές τιμές να αντιστοιχούν κατά το δυνατόν σε πιθανότητες ορθότητας των προβλέψεων κάθε συστήματος. Ο Freitag [48] προτείνει μια μέθοδο μετασχηματισμού βαθμών εμπιστοσύνης σε πιθανότητες ορθότητας, χρησιμοποιώντας *παλινδρόμηση (regression)* και περιγράφεται ως εξής:

Δοθέντος ενός αλγορίθμου L^i σχεδιασμένου για εξαγωγή πληροφορίας, η σχέση μεταξύ βαθμών εμπιστοσύνης και πιθανοτήτων ορθότητας στις προβλέψεις του αντίστοιχου εκπαιδευμένου συστήματος E^i, πάντα για μια συγκεκριμένη θεματική περιοχή, μπορεί να μοντελοποιηθεί με τη χρήση Q συναρτήσεων γραμμικής παλινδρόμησης, μιας για κάθε σχετικό πεδίο $f^1 ... f^Q$. Οι νέες συναρτήσεις εκτιμώνται μέσω *διασταυρωμένης επικύρωσης* (διαφορετικής από την αντίστοιχη για τη δημιουργία των διανυσμάτων χαρακτηριστικών σε μετα-επίπεδο) τριών βημάτων, ως εξής:

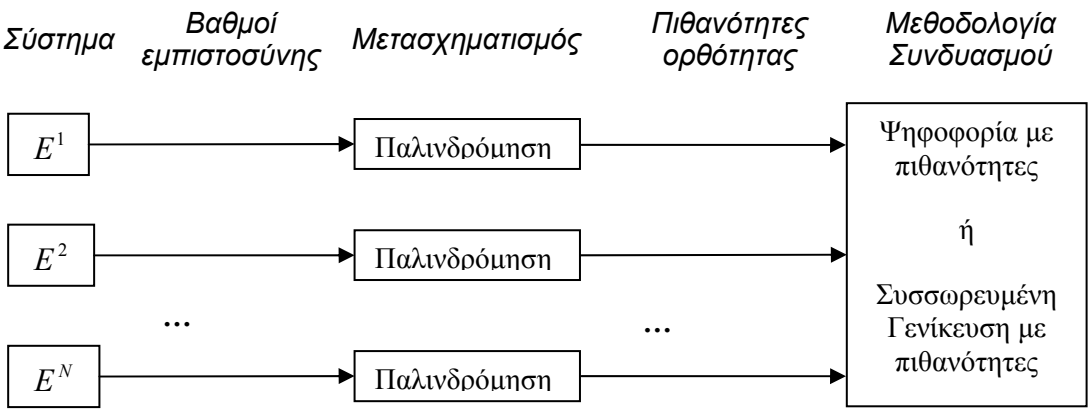
Το αρχικό σύνολο D των κειμένων εκπαίδευσης χωρίζεται σε τρία τμήματα περίπου ιδίου μεγέθους. Σε καθένα *j*-βήμα, *j=1…3*, της διαδικασίας διασταυρωμένης επικύρωσης, ο αλγόριθμος L^i εφαρμόζεται στο υποσύνολο $D \setminus D^j$ των κειμένων

εκπαίδευσης, και το εκπαιδευμένο σύστημα εφαρμόζεται στο σύνολο D^j των κειμένων επαλήθευσης. Όλα τα σχετικά παραδείγματα $< t(s,e), f^k, c^k >$ που έχουν αναγνωριστεί από το σύστημα σε όλα τα κείμενα επαλήθευσης του συνόλου D^j, όπου c^k ο βαθμός εμπιστοσύνης για το προβλεπόμενο πεδίο f^k, συγκεντρώνονται σε μια αρχική «αποθήκη» παραδειγμάτων. Στο τέλος ολόκληρης της διαδικασίας διασταυρωμένης επικύρωσης, τα παρακάτω βήματα λαμβάνουν χώρα για κάθε σχετικό πεδίο f:

1. Το διάστημα ανάμεσα στο μεγαλύτερο και το μικρότερο βαθμό εμπιστοσύνης μεταξύ των παραδειγμάτων που έχουν αναγνωριστεί ως σχετικά για το πεδίο f, χωρίζεται σε 10 ισομεγέθη διαστήματα. Τα σχετικά παραδείγματα ταξινομούνται στη συνέχεια σε αυτά τα 10 διαστήματα, ανάλογα με τον βαθμό εμπιστοσύνης που έχει καταχωρηθεί σε κάθε παράδειγμα.
2. Τα 10 διαστήματα που έχουν δημιουργηθεί χρησιμοποιούνται για την κατασκευή αντίστοιχων ζευγαριών (x, y), όπου x είναι το ενδιάμεσο του διαστήματος και y είναι η *ακρίβεια* των παραδειγμάτων που ανήκουν στο συγκεκριμένο διάστημα τιμών, δηλαδή το πηλίκο των σωστών παραδειγμάτων που έχουν αναγνωριστεί, προς το σύνολο των σωστών παραδειγμάτων για το πεδίο.
3. Τα 10 ζευγάρια χρησιμοποιούνται για τον υπολογισμό μιας συνάρτησης γραμμικής παλινδρόμησης για το πεδίο f.

Η παραπάνω διαδικασία επαναλαμβάνεται για όλα τα πεδία μιας θεματικής περιοχής και για όλους τους αλγορίθμους L^i, $i = 1...N.$ Χρησιμοποιώντας τις συναρτήσεις που έχουν υπολογιστεί, για κάθε παράδειγμα που αναγνωρίζεται ως σχετικό από ένα εκπαιδευμένο σύστημα κατά τη διαδικασία επαλήθευσης, εκχωρείται σε αυτό μια νέα τιμή που αντιπροσωπεύει την πιθανότητα το συγκεκριμένο παράδειγμα να είναι σωστό.

Το Σχήμα 5.3 δείχνει σχηματικά τον τρόπο συνδυασμού συστημάτων, υπό το πρίσμα της χρήσης βαθμών εμπιστοσύνης που έχουν μετατραπεί σε πιθανότητες ορθότητας. Πρέπει να σημειωθεί ότι η μεθοδολογία μετασχηματισμού των βαθμών εμπιστοσύνης στην έξοδο ενός συστήματος εξαγωγής πληροφορίας σε πιθανότητες ορθότητας, όπως περιγράφεται σε αυτή την ενότητα, δεν είναι μοναδική. Στην εργασία που περιγράφεται στην εργασία [62], το τμήμα της παλινδρόμησης παραλείπεται, ενώ η υπόλοιπη διαδικασία παραμένει η ίδια όπως περιγράφεται σε αυτή την ενότητα. Στη διατριβή αυτή διατηρείται το βήμα της παλινδρόμησης για λόγους σύγκρισης με την πολυστρατηγική μάθηση για εξαγωγή πληροφορίας [48].



Σχήμα 5.3 Σχηματική αναπαράσταση του συνδυασμού συστημάτων εξαγωγής πληροφορίας, υπό το πρίσμα της χρήσης πιθανοτήτων ορθότητας στην έξοδο των συστημάτων.

Πρέπει επίσης να σημειωθεί ότι εναλλακτικοί τύποι παλινδρόμησης, πλην της γραμμικής που αναφέρεται σε αυτή την ενότητα, μπορούν να χρησιμοποιηθούν. Στα πλαίσια αυτής της διατριβής εξετάστηκαν οι περιπτώσεις της *τοπικά σταθμικής παλινδρόμησης (locally weighted regression)*, *δέντρων παλινδρόμησης (regression trees)*, και των *μοντέλων δέντρων (model trees)*. Τα αποτελέσματα δεν ήταν στατιστικά πιο σημαντικά από αυτά της γραμμικής παλινδρόμησης, η οποία διατηρήθηκε για την πραγματοποίηση αξιόπιστης σύγκρισης με την πολυστρατηγική μάθηση [48].

5.7 Αξιολόγηση συσσωρευμένης γενίκευσης

Σε αυτή την ενότητα παρουσιάζονται τα αποτελέσματα αξιολόγησης της συσσωρευμένης γενίκευσης στις πέντε θεματικές περιοχές ενδιαφέροντος, συγκρίνοντας ταυτόχρονα με τα καλύτερα αποτελέσματα που επιτεύχθηκαν σε βασικό επίπεδο για να διαπιστωθεί εάν επιτυγχάνεται βελτίωση σε μετα-επίπεδο. Η σύγκριση συσσωρευμένης γενίκευσης και ψηφοφορίας παρουσιάζεται στο Κεφάλαιο 6, μαζί με μια ολοκληρωμένη ανάλυση της απόδοσης των δύο μεθοδολογιών.

5.7.1 Παρουσίαση αναλυτικών αποτελεσμάτων

Οι Πίνακες 5.3 έως 5.7 παρουσιάζουν τα αποτελέσματα που επιτεύχθηκαν σε μετα-επίπεδο από τους ταξινομητές που αξιολογήθηκαν, τόσο με χρήση ονομαστικών τιμών στην αναπαράσταση των διανυσμάτων χαρακτηριστικών όσο και με χρήση πιθανοτήτων. Οι θετικές τιμές που φαίνονται σε πλάγια γραφή δίπλα στις τιμές για το $F1$ δείχνουν βελτίωση για το $F1$ κατά το χειρισμό των αγνοούμενων τιμών στην αναπαράσταση των διανυσμάτων χαρακτηριστικών του μετα-επιπέδου. Οι αρνητικές τιμές δείχνουν το αντίθετο, δηλαδή χειροτέρευση για το $F1$.

Πίνακας 5.3 Αποτελέσματα συσσωρευμένης γενίκευσης για τα πανεπιστημιακά μαθήματα της επιστήμης υπολογιστών.

Ταξινομητής	Συσσωρευμένη γενίκευση με χρήση ονομαστικών τιμών				Συσσωρευμένη γενίκευση με χρήση πιθανοτήτων			
%	*P*	*R*	$F1$		*P*	*R*	$F1$	
IB1	72.10	60.90	66.03	*0.00*	75.41	66.84	70.87	*0.07*
J48	75.10	38.37	50.79	*12.7*	74.65	66.32	70.24	*-0.27*
LogitBoost	81.32	52.66	63.92	*1.76*	79.03	66.01	71.93	*-0.84*
MLR	70.88	44.42	54.62	*9.63*	83.88	60.79	70.50	*0.20*
NaiveBayes	74.54	50.37	60.11	*5.20*	58.38	73.72	65.16	*0.91*
SMO	67.05	49.22	56.76	*5.92*	78.46	60.38	68.24	*-0.39*

Πίνακας 5.4 Αποτελέσματα συσσωρευμένης γενίκευσης για τα ερευνητικά προγράμματα.

Ταξινομητής	Συσσωρευμένη γενίκευση με χρήση ονομαστικών τιμών				Συσσωρευμένη γενίκευση με χρήση πιθανοτήτων			
%	*P*	*R*	$F1$		*P*	*R*	$F1$	
IB1	50.36	77.91	61.18	*0.00*	69.32	64.04	66.58	*0.00*
J48	54.00	56.78	55.36	*3.39*	75.78	67.52	71.41	*-0.01*
LogitBoost	68.84	63.59	66.05	*-5.07*	78.21	64.45	70.67	*-0.20*
MLR	54.28	69.34	60.89	*-0.12*	73.45	58.60	65.19	*0.00*
NaiveBayes	54.41	69.39	60.99	*0.15*	59.75	75.74	66.80	*4.49*
SMO	50.18	63.29	55.98	*3.17*	70.44	62.73	66.36	*0.33*

Πίνακας 5.5 Αποτελέσματα συσσωρευμένης γενίκευσης για τους φορητούς υπολογιστές.

Ταξινομητής	Συσσωρευμένη γενίκευση με χρήση ονομαστικών τιμών				Συσσωρευμένη γενίκευση με χρήση πιθανοτήτων			
	P	*R*	$F1$		*P*	*R*	$F1$	
IB1	58.09	72.08	64.43	*0.00*	78.59	62.25	69.48	*-0.15*
J48	77.00	52.68	62.56	*3.12*	80.56	62.93	70.66	*-0.05*
LogitBoost	79.52	60.10	68.46	*-3.12*	84.54	62.04	71.56	*0.67*
MLR	74.40	59.68	66.23	*0.48*	86.53	58.79	70.02	*0.16*
NaiveBayes	69.27	65.42	67.29	*-1.70*	52.24	74.40	61.38	*0.07*
SMO	76.86	58.84	66.65	*0.75*	84.98	60.86	70.02	*1.07*

Πίνακας 5.6 Αποτελέσματα συσσωρευμένης γενίκευσης για τις αγγελίες εργασίας.

Ταξινομητής	Συσσωρευμένη γενίκευση με χρήση ονομαστικών τιμών				Συσσωρευμένη γενίκευση με χρήση πιθανοτήτων			
	P	*R*	$F1$		*P*	*R*	$F1$	
IB1	73.14	89.69	80.58	*0.00*	86.48	81.44	83.88	*-0.12*
J48	86.72	81.32	83.93	*-0.50*	89.53	81.31	85.22	*0.10*
LogitBoost	89.89	81.82	85.67	*-3.30*	90.27	82.00	85.94	*-0.04*
MLR	84.51	84.71	84.61	*-2.09*	90.92	79.71	84.95	*0.04*
NaiveBayes	76.94	87.24	81.77	*-0.67*	68.98	87.73	77.23	*-0.03*
SMO	88.11	81.22	84.52	*0.92*	89.43	80.54	84.75	*0.25*

Πίνακας 5.7 Αποτελέσματα συσσωρευμένης γενίκευσης για τις ανακοινώσεις σεμιναρίων.

Ταξινομητής	Συσσωρευμένη γενίκευση με χρήση ονομαστικών τιμών				Συσσωρευμένη γενίκευση με χρήση πιθανοτήτων			
	P	*R*	$F1$		*P*	*R*	$F1$	
IB1	87.42	87.25	87.34	*0.00*	93.27	84.11	88.45	*-0.01*
J48	89.50	85.83	87.63	*-0.06*	94.66	85.17	89.66	*-0.03*
LogitBoost	92.56	84.74	88.48	*0.12*	94.69	85.80	90.03	*0.14*
MLR	89.59	86.89	88.22	*0.30*	93.38	84.62	88.78	*-0.16*
NaiveBayes	87.15	86.97	87.06	*0.15*	89.58	87.42	88.49	*0.15*
SMO	92.03	84.46	88.09	*0.08*	93.12	84.89	88.82	*0.07*

Όσον αφορά το χειρισμό των αγνοούμενων τιμών στις αναπαραστάσεις των διανυσμάτων χαρακτηριστικών του μετα-επιπέδου, οι Πίνακες 5.3 έως 5.7 δείχνουν βελτίωση σε αρκετές περιπτώσεις με σημαντικότερη εκείνη του ταξινομητή *J48* με χρήση ονομαστικών τιμών. Το τελικό συμπέρασμα, όμως, είναι ότι ο χειρισμός αγνοούμενων τιμών δεν οδηγεί σε σημαντική βελτίωση τα καλύτερα αποτελέσματα που επιτυγχάνονται σε μετα-επίπεδο για κάθε θεματική περιοχή ενδιαφέροντος.

5.7.2 Σύγκριση με το βασικό επίπεδο

Ο Πίνακας 5.8 συνοψίζει από τους Πίνακες 5.3 έως 5.7 τις καλύτερες τιμές για το $F1$ ανά θεματική περιοχή, τόσο για τη συσσωρευμένη γενίκευση με χρήση ονομαστικών τιμών, όσο και για την περίπτωση της χρήσης πιθανοτήτων. Οι Πίνακες 5.9 έως 5.11 συγκρίνουν τις δύο προσεγγίσεις συσσωρευμένη γενίκευσης για εξαγωγή πληροφορίας με το καλύτερο σύστημα του βασικού επιπέδου. Η σύγκριση πραγματοποιείται με βάση τις στατιστικά πιο σημαντικές *νίκες* έναντι *ηττών* στις πέντε θεματικές περιοχές ενδιαφέροντος, χρησιμοποιώντας τις μετρικές *ακρίβεια*, *ανάκληση* και $F1$ αντίστοιχα. Υπενθυμίζεται ότι η μέτρηση της στατιστικής σημαντικότητας πραγματοποιείται με βάση το γνωστό τεστ paired t-test [38] με ποσοστό σημαντικότητας το 95%.

Πίνακας 5.8 Τα καλύτερα αποτελέσματα που έχουν επιτευχθεί από τη συσσωρευμένη γενίκευση για τις πέντε θεματικές περιοχές ενδιαφέροντος καθώς και η σύγκρισή τους με τα καλύτερα αποτελέσματα του βασικού επιπέδου.

	Εξαγωγή πληροφορίας σε βασικό επίπεδο			Συσσώρευση με χρήση ονομαστικών τιμών			Συσσώρευση με χρήση πιθανοτικών τιμών		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Μαθήματα	71.39	60.90	65.73	72.10	60.90	66.03	79.03	66.01	71.93
Προγράμματα	56.24	68.18	61.64	68.84	63.59	66.05	75.78	67.52	71.41
Φορητοί	62.29	65.42	63.81	79.52	60.10	68.46	85.03	62.76	72.23
Αγγελίες	87.70	79.18	83.22	89.89	81.82	85.67	90.27	82.00	85.94
Ανακοινώσεις	91.39	81.63	86.23	92.56	84.74	88.48	94.69	85.80	90.03

Πίνακας 5.9 Στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, χρησιμοποιώντας τη μετρική *ακρίβεια*, στις πέντε θεματικές περιοχές ενδιαφέροντος.

	Βασικό επίπεδο	Συσσώρευση απλή	Συσσώρευση με πιθανότητες
Βασικό επίπεδο		0\5	0\5
Συσσώρευση απλή	5\0		0\3
Συσσώρευση με πιθανότητες	5\0	3\0	

Πίνακας 5.10 Στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, χρησιμοποιώντας τη μετρική *ανάκληση*, στις πέντε θεματικές περιοχές ενδιαφέροντος.

	Βασικό επίπεδο	Συσσώρευση απλή	Συσσώρευση με πιθανότητες
Βασικό επίπεδο		2\2	2\3
Συσσώρευση απλή	2\2		0\3
Συσσώρευση με πιθανότητες	3\2	3\0	

Πίνακας 5.11 Στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, χρησιμοποιώντας τη μετρική $F1$, στις πέντε θεματικές περιοχές ενδιαφέροντος.

	Βασικό επίπεδο	Συσσώρευση απλή	Συσσώρευση με πιθανότητες
Βασικό επίπεδο		0\2	0\5
Συσσώρευση απλή	2\0		0\4
Συσσώρευση με πιθανότητες	5\0	4\0	

Οι Πίνακες 5.8 έως 5.11 δείχνουν ότι η συσσωρευμένη γενίκευση με χρήση πιθανοτήτων επιτυγχάνει καλύτερο $F1$ από το καλύτερο σύστημα του βασικού επιπέδου για κάθε θεματική περιοχή. Η συσσωρευμένη γενίκευση με χρήση ονομαστικών τιμών, από την άλλη πλευρά, επιτυγχάνει στατιστικά καλύτερο $F1$ σε μετα-επίπεδο μόνο για δύο περιοχές. Η μεγάλη βελτίωση που επιτυγχάνει η απλή συσσωρευμένη γενίκευση στα ερευνητικά προγράμματα και στους φορητούς υπολογιστές δεν είναι συνεπής σε όλα τα βήματα της διασταυρωμένης επικύρωσης κατά τη διαδικασία αξιολόγησης και γι' αυτό μετρήθηκε ως στατιστικά ασήμαντη.

Η χρήση πιθανοτήτων στη συσσωρευμένη γενίκευση επιτυγχάνει καλύτερο $F1$ σε μετα-επίπεδο σε σχέση με τη χρήση ονομαστικών τιμών σε τέσσερις από τις πέντε περιοχές. Κάτι τέτοιο είναι αναμενόμενο, λόγω της επιπρόσθετης πληροφορίας (πιθανότητες ορθότητας στις προβλέψεις κάθε συστήματος του βασικού επιπέδου) που είναι διαθέσιμη σε μετα-επίπεδο και μπορεί να εκμεταλλευτεί ένας ταξινομητής. Μόνο για την περιοχή των αγγελιών εργασίας, η διαφορά μετρήθηκε ως στατιστικά μη σημαντική.

Επίσης, η συσσωρευμένη γενίκευση με χρήση πιθανοτήτων βελτιώνει την ακρίβεια σε σχέση με το καλύτερο σύστημα του βασικού επιπέδου και στις πέντε θεματικές

περιοχές. Η βελτίωση είναι πιο εντυπωσιακή για τις περιοχές των ερευνητικών προγραμμάτων και φορητών υπολογιστών. Επίσης και η απλή συσσώρευση με χρήση ονομαστικών τιμών βελτιώνει την ακρίβεια σε μετα-επίπεδο για κάθε περιοχή. Όμως η χρήση πιθανοτήτων στη συσσωρευμένη γενίκευση επιτυγχάνει μεγαλύτερη ακρίβεια σε μετα-επίπεδο σε τρεις από τις πέντε περιοχές, σε σχέση με τη χρήση ονομαστικών τιμών, ενώ στις υπόλοιπες δύο η διαφορά μετρήθηκε ως στατιστικά ασήμαντη.

Στο θέμα της ανάκλησης, η κατάσταση δεν είναι ξεκάθαρη. Η συσσωρευμένη γενίκευση με χρήση πιθανοτήτων βελτιώνει την ανάκληση σε σχέση με το βασικό επίπεδο στις τρεις από τις πέντε περιοχές και συγκεκριμένα στα μαθήματα υπολογιστών, στις αγγελίες εργασίας και στις ανακοινώσεις σεμιναρίων, ενώ στις άλλες δύο η ανάκληση βλάπτεται σε μετα-επίπεδο. Η βελτίωση στην ακρίβεια, όμως, είναι πολύ πιο εντυπωσιακή, οδηγώντας σε σημαντική αύξηση του $F1$ σε μετα-επίπεδο. Η συσσώρευση με χρήση ονομαστικών τιμών βελτιώνει την ανάκληση σε δύο περιοχές ενώ την βλάπτει σε άλλες δύο. Τέλος, η χρήση πιθανοτήτων ορθότητας στη συσσωρευμένη γενίκευση οδηγεί σε καλύτερη ανάκληση, σε σχέση με τη χρήση ονομαστικών τιμών, σε τρεις από τις πέντε περιοχές, ενώ η διαφορά είναι στατιστικά ασήμαντη στις υπόλοιπες δύο θεματικές περιοχές.

5.7.3 Σύγκριση ταξινομητών σε μετα-επίπεδο με χρήση πιθανοτήτων

Ο Πίνακας 5.12 συγκρίνει όλους τους ταξινομητές της χρήσης συσσωρευμένης γενίκευσης με χρήση πιθανοτήτων, που απέδωσε τα καλύτερα αποτελέσματα εξαγωγής σε μετα-επίπεδο. Η σύγκριση πραγματοποιείται με βάση τις στατιστικά καλύτερες *νίκες* έναντι *ηττών*, χρησιμοποιώντας το $F1$, στις πέντε θεματικές περιοχές ενδιαφέροντος.

Πίνακας 5.12 Σύγκριση ταξινομητών σε μετα-επίπεδο, για τη συσσώρευση με πιθανότητες, με βάση τις στατιστικά σημαντικότερες *νίκες* έναντι ηττών, χρησιμοποιώντας το $F1$, στις πέντε θεματικές περιοχές ενδιαφέροντος.

	IB1	J48	LogitBoost	MLR	NaiveBayes	SMO
IB1		0\3	0\5	1\1	3\0	0\2
j48	3\0		0\3	2\0	5\0	1\0
LogitBoost	5\0	3\0		5\0	5\0	5\0
MLR	1\1	0\2	0\5		3\1	0\2
NaiveBayes	0\3	0\5	0\5	1\3		0\2
SMO	2\0	0\1	0\5	2\0	2\0	

Ο Πίνακας 5.12 δείχνει ότι δεν υπάρχει κανένας καθολικά καλύτερος ταξινομητής σε μετα-επίπεδο σε όλες τις περιοχές. Συγκεκριμένοι ταξινομητές ταιριάζουν καλύτερα στα χαρακτηριστικά κάποιων συλλογών, οδηγώντας σε καλύτερα αποτελέσματα από

άλλους. Ο Πίνακας 5.12 δείχνει παρόλα αυτά την ανωτερότητα δύο ταξινομητών: του *LogitBoost* (κυρίως) και του *j48*. Πληθώρα άλλων ταξινομητών μπορούν επίσης να αξιολογηθούν σε μετα-επίπεδο. Ο σκοπός των πειραμάτων που πραγματοποιήθηκαν σε αυτή τη διατριβή είναι να αναδειχτεί η αποτελεσματικότητα της προτεινόμενης μεθοδολογίας συσσωρευμένης γενίκευσης για το συνδυασμό συστημάτων εξαγωγής πληροφορίας. Οι περισσότεροι από τους ταξινομητές που χρησιμοποιήθηκαν έχουν επίσης χρησιμοποιηθεί για αξιολόγηση σε πρόσφατες εργασίες που αφορούν τη συσσωρευμένη γενίκευση για προβλήματα ταξινόμησης [41, 98, 107]. Στις εργασίες αυτές η χρήση του *MLR* ταξινομητή αποδείχτηκε ιδιαίτερα αποτελεσματική. Αντίθετα, κάτι τέτοιο δεν αποδείχτηκε στην περίπτωση της εξαγωγής πληροφορίας. Θα πρέπει να υπενθυμίσουμε, βέβαια, ότι στις προαναφερθείσες εργασίες, κάθε ταξινομητής του βασικού επιπέδου εξάγει ένα διάνυσμα χαρακτηριστικών που αντιστοιχεί σε μια πιθανοτική κατανομή σε όλες τις σχετικές κλάσεις. Κάτι τέτοιο όμως δεν ισχύει στην περίπτωση των συστημάτων εξαγωγής πληροφορίας. Από την άλλη πλευρά, τα αποτελέσματα δείχνουν ότι στην περίπτωση της εξαγωγής πληροφορίας, ο ταξινομητής *LogitBoost*, αποδεικνύεται ιδιαίτερα αποτελεσματικός σε μετα-επίπεδο.

5.7.4 Αξιολόγηση σε ζευγάρια συστημάτων

Πειράματα πραγματοποιήθηκαν επίσης και σε όλους τους δυνατούς συνδυασμούς ζευγαριών συστημάτων εξαγωγής πληροφορίας. Στόχος ήταν να διερευνηθεί εάν ο συνδυασμός, μέσω συσσωρευμένης γενίκευσης, και των τριών συστημάτων του βασικού επιπέδου επιτυγχάνει καλύτερα αποτελέσματα εξαγωγής από το συνδυασμό ζευγαριών συστημάτων. Ο Πίνακας 5.13 συγκρίνει όλους τους συνδυασμούς συστημάτων του βασικού επιπέδου, μέσω συσσωρευμένης γενίκευσης με πιθανότητες. Η σύγκριση πραγματοποιείται με βάση τις στατιστικά σημαντικότερες *νίκες* στις πέντε θεματικές περιοχές, χρησιμοποιώντας το $F1$. Για τη διεξαγωγή δίκαιων συγκρίσεων, τα αποτελέσματα του ίδιου ταξινομητή *(LogitBoost)* χρησιμοποιούνται για όλους τους συνδυασμούς συστημάτων. Το Παράρτημα Α.7 δείχνει αναλυτικές τιμές και στις τρεις μετρικές αξιολόγησης για όλους τους συνδυασμούς συστημάτων.

Το γεγονός ότι ο συνδυασμός περισσοτέρων συστημάτων οδηγεί σε καλύτερα αποτελέσματα δεν αποτελεί έκπληξη. Όμως ο συνδυασμός των HMMs με $(LP)^2$ σε μια περιοχή (μαθήματα επιστήμης υπολογιστών) οδηγεί σε ισοδύναμα αποτελέσματα με το συνδυασμό και των τριών συστημάτων, οδηγώντας στο συμπέρασμα ότι η συνεισφορά του BWI κατά το συνδυασμό δεν είναι σημαντική για τη συγκεκριμένη περιοχή.

Πίνακας 5.13 Στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, χρησιμοποιώντας το $F1$, του συνδυασμού -μέσω συσσωρευμένης γενίκευσης με πιθανότητες- των συστημάτων του βασικού επιπέδου της γραμμής, έναντι του αντίστοιχου της στήλης.

	BWI +HMM	BWI+$(LP)^2$	HMM+$(LP)^2$	BWI+HMM+$(LP)^2$
BWI +HMM		0\4	0\5	0\5
BWI+$(LP)^2$	4\0		0\4	0\5
HMM+$(LP)^2$	5\0	4\0		0\4
BWI+HMM+$(LP)^2$	5\0	5\0	4\0	

Επιπλέον, ο συνδυασμός μέσω συσσωρευμένης γενίκευσης των HMMs με $(LP)^2$ αποδεικνύεται καλύτερος από τον αντίστοιχο συνδυασμό των BWI με $(LP)^2$, το οποίο δε δικαιολογείται πάντα από την απόδοση ξεχωριστά των συστημάτων του βασικού επιπέδου. Για παράδειγμα, το σύστημα του BWI επιτυγχάνει καλύτερο $F1$ από το αντίστοιχο των HMMs στις αγγελίες εργασίας και στις ανακοινώσεις σεμιναρίων. Ο Πίνακας 3.11 δικαιολογεί αυτή τη συμπεριφορά, παρατηρώντας μεγαλύτερη συσχέτιση ανάμεσα στα συστήματα BWI και HMMs στις αγγελίες εργασίας και ανακοινώσεις σεμιναρίων, από ότι ανάμεσα στα συστήματα HMMs και $(LP)^2$.

Τα αποτελέσματα στους Πίνακες 3.2 έως 3.6 δείχνουν ότι το σύστημα του BWI υποφέρει από χαμηλή ανάκληση στις περισσότερες περιοχές, σε σύγκριση με τα υπόλοιπα δύο συστήματα του βασικού επιπέδου. Μια γενική οδηγία, επομένως, για την επιλογή των συστημάτων εκείνων τα οποία θέλουμε να συνδυάσουμε μέσω συσσωρευμένης γενίκευσης, θα μπορούσε να είναι η προτίμηση προς τα συστήματα εκείνα τα οποία επιτυγχάνουν καλύτερη ανάκληση. Η μόνη εξαίρεση αφορά τα ερευνητικά προγράμματα, όπου το σύστημα του $(LP)^2$ επιτυγχάνει σημαντικά χαμηλότερη ανάκληση από το αντίστοιχο των HMMs, αλλά η συσσώρευση με πιθανότητες των συστημάτων BWI και $(LP)^2$ επιτυγχάνει καλύτερη απόδοση από την αντίστοιχη των συστημάτων BWI και HMMs. Ο Πίνακας 3.11 δικαιολογεί επίσης αυτή τη συμπεριφορά, όπου παρατηρείται μεγαλύτερος βαθμός συσχέτισης ανάμεσα στα συστήματα BWI και HMMs, από ότι μεταξύ των συστημάτων BWI και $(LP)^2$.

5.8 Συμπεράσματα

Η χρήση συσσωρευμένης γενίκευσης με απλές ονομαστικές τιμές οδήγησε σε καλύτερα αποτελέσματα στην εξαγωγή πληροφορίας σε μετα-επίπεδο μόνο στις δύο από τις πέντε θεματικές περιοχές ενδιαφέροντος. Από την άλλη πλευρά, η χρήση συσσωρευμένης γενίκευσης με πιθανότητες αποδείχτηκε καθολικά καλύτερη από όλα τα συστήματα του βασικού επιπέδου και στις πέντε θεματικές περιοχές. Οδηγεί πάντα, επίσης, σε πιο ακριβή αποτελέσματα στην εξαγωγή πληροφορίας σε μετα-επίπεδο σε

σχέση με το καλύτερο σύστημα του βασικού επιπέδου, ενώ η ανάκληση βελτιώνεται στις περισσότερες περιοχές. Συγκρίνοντας τις δύο τεχνικές, η συσσωρευμένη γενίκευση με πιθανότητες επιτυγχάνει καλύτερα αποτελέσματα από τη συσσωρευμένη γενίκευση με ονομαστικές τιμές σε όλες περιοχές, ενώ μόνο σε μια (στις αγγελίες εργασίας) η διαφορά στα αποτελέσματα μετρήθηκε ως στατιστικά μη σημαντική.

Η συσσωρευμένη γενίκευση ζευγών συστημάτων του βασικού επιπέδου δεν έδωσε καλύτερα αποτελέσματα στην εξαγωγή πληροφορίας από ότι ο αντίστοιχος συνδυασμός και των τριών διαθέσιμων συστημάτων. Το συμπέρασμα αυτό ήταν αναμενόμενο. Η εξήγηση των αποτελεσμάτων, όμως, οδήγησε στο ιδιαίτερο χρήσιμο συμπέρασμα ότι για το συνδυασμό θα πρέπει να προτιμώνται συστήματα εξαγωγής πληροφορίας με πολύ καλή *ανάκληση* στις προβλέψεις τους, καθώς η *ακρίβεια* βελτιώνεται πάντα από τη συσσωρευμένη γενίκευση σε μετα-επίπεδο.

Οι πιθανότητες που χρησιμοποιούνται κατά τη συσσωρευμένη γενίκευση είναι οι ίδιες με αυτές που χρησιμοποιούνται και κατά την πιθανοτική ψηφοφορία και αντιστοιχούν σε πιθανότητες ορθότητας στην έξοδο των συστημάτων εξαγωγής του βασικού επιπέδου. Αυτό που απομένει ακόμα είναι μια συγκριτική αξιολόγηση όλων των τεχνικών συνδυασμού (ψηφοφορίας και συσσωρευμένης γενίκευσης) που προτείνονται σε αυτή τη διατριβή. Χρειάζεται επίσης μια περαιτέρω ανάλυση των αποτελεσμάτων ώστε να γίνουν όσο το δυνατόν κατανοητά τα χαρακτηριστικά της συμπεριφοράς όλων των προτεινόμενων τεχνικών συνδυασμού για την εξαγωγή πληροφορίας.

ΚΕΦΑΛΑΙΟ 6

ΣΥΓΚΡΙΣΗ ΜΕΘΟΔΩΝ ΚΑΙ ΑΝΑΛΥΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Η αξιολόγηση που έχει γίνει μέχρι τώρα έχει αναδείξει την αποτελεσματικότητα των τεχνικών ψηφοφορίας και συσσωρευμένης γενίκευσης για το συνδυασμό συστημάτων εξαγωγής πληροφορίας. Στην Ενότητα 6.1 συγκρίνονται αναλυτικά οι τεχνικές ψηφοφορίας και συσσωρευμένης γενίκευσης, τόσο με βάση τα αποτελέσματα που επιτεύχθηκαν, όσο και με βάση το υπολογιστικό κόστος τους. Στόχος είναι να διαπιστωθεί εάν η συσσωρευμένη γενίκευση οδηγεί σε καλύτερα αποτελέσματα εξαγωγής σε μετα-επίπεδο από ότι η ψηφοφορία κι αν δικαιολογεί επομένως το επιπρόσθετο υπολογιστικό κόστος που έχει σε σχέση με την ψηφοφορία.

Στην Ενότητα 6.2 επιχειρείται μια ανάλυση των αποτελεσμάτων που επιτεύχθηκαν από όλες τις τεχνικές συνδυασμού, με βάση τη διαφορετικότητα στις προβλέψεις των συστημάτων του βασικού επιπέδου. Στόχος είναι να διερευνηθεί η συμπεριφορά της τόσο της συσσωρευμένης γενίκευσης όσο και της ψηφοφορίας, με βάση τη διαφορετικότητα αυτή. Ανάλογη ανάλυση απουσιάζει μέχρι τώρα από τη διεθνή βιβλιογραφία. Τέλος, η Ενότητα 6.3 συνοψίζει τα συμπεράσματα του κεφαλαίου αυτού.

6.1 Σύγκριση ψηφοφορίας με συσσωρευμένη γενίκευση

Η σύγκριση συσσωρευμένης γενίκευσης και ψηφοφορίας είναι απαραίτητη με κύριο στόχο να διαπιστωθεί εάν η πρώτη αξίζει το επιπρόσθετο υπολογιστικό κόστος.

6.1.1 Σύγκριση με βάση την συνολική απόδοση

Ο Πίνακας 6.1 δείχνει τις καλύτερες τιμές για το $F1$ που επιτυγχάνονται από όλες τις τεχνικές συνδυασμού (ψηφοφορίας και συσσωρευμένης γενίκευσης) στις πέντε θεματικές περιοχές ενδιαφέροντος, καθώς και από το καλύτερο σύστημα του βασικού επιπέδου. Οι Πίνακες 6.2 έως 6.4 συγκρίνουν τις τεχνικές συνδυασμού με βάση τις στατιστικά σημαντικότερες *νίκες* έναντι *ηττών* στις πέντε θεματικές περιοχές, χρησιμοποιώντας και τις τρεις μετρικές αξιολόγησης (ακρίβεια, ανάκληση, $F1$) για τους Πίνακες 6.2 έως 6.4 αντίστοιχα. Για την επίτευξη αντικειμενικών συγκρίσεων, τα αποτελέσματα ενός μόνο ταξινομητή *(LogitBoost)* χρησιμοποιήθηκαν για τη συσσωρευμένη γενίκευση.

Πίνακας 6.1 Καλύτερες τιμές για το $F1$ από όλες τις τεχνικές συνδυασμού, καθώς και από το καλύτερο σύστημα του βασικού επιπέδου για κάθε περιοχή.

	Βασικό	MVotM	MVotF	PVotM	PVotF	Συσ/ση απλή	Συσ/ση πιθαν.
Μαθήματα	65.73	65.59	60.29	65.65	70.64	63.92	71.93
Προγράμματα	61.64	60.71	67.39	60.75	65.75	66.05	70.66
Φορητοί	63.81	62.37	67.60	62.76	71.03	68.46	71.55
Αγγελίες	83.22	79.90	83.85	79.99	83.15	85.67	85.94
Σεμινάρια	86.23	86.87	87.13	86.90	88.02	88.48	90.03

Πίνακας 6.2 Κάθε κελί δείχνει τις στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, με βάση την *ακρίβεια*, στις πέντε θεματικές περιοχές, του συστήματος εξαγωγής πληροφορίας στη γραμμή, έναντι του αντιστοίχου της στήλης.

	Βασικό	MVotM	MVotF	PVotM	PVotF	Συσ/ση απλή	Συς/ση πιθαν.
Βασικό		5\0	0\5	5\0	2\2	0\5	0\5
MVotM	0\5		0\5	0\1	0\5	0\5	0\5
MVotF	5\0	5\0		5\0	5\0	2\0	3\1
PVotM	0\5	1\0	0\5		0\5	0\5	0\5
PVotF	2\2	5\0	0\5	5\0		0\5	0\5
Συσ/ση απλή	5\0	5\0	0\2	5\0	5\0		0\3
Συσ/ση πιθαν.	5\0	5\0	1\3	5\0	5\0	3\0	

Πίνακας 6.3 Κάθε κελί δείχνει τις στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, με βάση την *ανάκληση*, στις πέντε θεματικές περιοχές, του συστήματος εξαγωγής πληροφορίας στη γραμμή, έναντι του αντιστοίχου της στήλης.

	Βασικό	MVotM	MVotF	PVotM	PVotF	Συσ/ση απλή	Συσ/ση πιθαν.
Βασικό		0\5	4\0	0\5	0\4	2\2	2\3
MVotM	5\0		5\0	0\1	5\0	5\0	5\0
MVotF	0\4	0\5		0\5	0\5	0\4	0\4
PVotM	5\0	1\0	5\0		5\0	5\0	5\0
PVotF	4\0	0\5	5\0	0\5		4\0	4\0
Συσ/ση απλή	2\2	0\5	4\0	0\5	0\4		0\3
Συσ/ση πιθαν.	3\2	0\5	4\0	0\5	0\4	3\0	

Πίνακας 6.4 Κάθε κελί δείχνει τις στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, με βάση τη μετρική $F1$, στις πέντε θεματικές περιοχές, του συστήματος εξαγωγής πληροφορίας στη γραμμή, έναντι του αντιστοίχου της στήλης.

	Βασικό	MVotM	MVotF	PVotM	PVotF	Συσ/ση απλή	Συσ/ση πιθαν.
Βασικό		2\0	1\2	1\0	0\4	0\2	0\5
MVotM	0\2		1\3	0\1	0\5	0\3	0\5
MVotF	2\1	3\1		3\1	1\3	0\3	0\5
PVotM	0\1	1\0	1\3		0\5	0\3	0\5
PVotF	4\0	5\0	3\1	5\0		2\2	0\3
Συσ/ση απλή	2\0	3\0	3\0	3\0	2\2		0\4
Συσ/ση πιθαν.	5\0	5\0	5\0	5\0	3\0	4\0	

Ο Πίνακας 6.2 δείχνει ότι η συσσωρευμένη γενίκευση, τόσο με χρήση ονομαστικών τιμών όσο και με χρήση πιθανοτήτων, επιτυγχάνει μεγαλύτερη ακρίβεια από το *PVotF* και στις 5 θεματικές περιοχές. Μόνο το σχήμα ψηφοφορίας *MVotF* επιτυγχάνει μεγαλύτερη ακρίβεια από τη συσσωρευμένη γενίκευση στις περισσότερες περιοχές, που συνοδεύεται όμως από πολύ μεγαλύτερη μείωση στην ανάκληση και σε χειρότερο τελικά $F1$. Συνολικά, η συσσωρευμένη γενίκευση με χρήση πιθανοτήτων επιτυγχάνει μεγαλύτερη ακρίβεια σε μετα-επίπεδο από το καλύτερο σύστημα το βασικού επιπέδου για κάθε περιοχή αλλά και από το καλύτερο σχήμα ψηφοφορίας, που είναι το *PVotF*.

Ο Πίνακας 6.3 δείχνει ότι στο θέμα της ανάκλησης, η ψηφοφορία τα πηγαίνει καλύτερα από ότι η συσσωρευμένη γενίκευση. Όπως έχει εξηγηθεί στο Κεφάλαιο 4, η ανάκληση για τα σχήματα ψηφοφορίας *MVotM* και *PVotM* αποτελεί μια αρκετά καλή προσέγγιση της μέγιστης ανάκλησης που μπορεί να επιτευχθεί σε μετα-επίπεδο και στις πέντε περιοχές. Η συσσωρευμένη γενίκευση απέχει από αυτές τις προσεγγιστικά μέγιστες τιμές, παρόλο που βελτιώνει, με χρήση πιθανοτήτων, την ανάκληση του καλύτερου συστήματος του βασικού επιπέδου σε τρεις από τις πέντε περιοχές.

Από τους Πίνακες 6.3 και 6.4 φαίνεται ότι η συσσωρευμένη γενίκευση με χρήση πιθανοτήτων επιτυγχάνει μεγαλύτερο $F1$ από το σχήμα ψηφοφορίας *PVotF* και στις πέντε θεματικές περιοχές. Μόνο σε δύο περιοχές (μαθήματα επιστήμης υπολογιστών και φορητοί υπολογιστές) οι διαφορές μεταξύ συσσωρευμένης γενίκευσης με πιθανότητες και *PVotF* μετρήθηκαν ως στατιστικά μη σημαντικές. Από την άλλη πλευρά, η συσσωρευμένη γενίκευση με χρήση ονομαστικών τιμών επιτυγχάνει καλύτερο $F1$ από το *PVotF* σε δύο θεματικές περιοχές, ενώ το *PVotF* υπερτερεί σε άλλες δύο.

Το τελικό συμπέρασμα είναι η συσσωρευμένη γενίκευση με χρήση πιθανοτήτων είναι η καλύτερη μέθοδος συνδυασμού συστημάτων εξαγωγής από όλες όσες αξιολογήθηκαν.

6.1.2 Σύγκριση με βάση τα πεδία

Ο Πίνακας 6.5 συγκρίνει όλες τις τεχνικές συνδυασμού με βάση τις στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, χρησιμοποιώντας το $F1$, σε όλα τα πεδία αθροιστικά (σύνολο 45) και στις πέντε περιοχές. Ο Πίνακας 6.5 δείχνει την ανωτερότητα της συσσωρευμένης γενίκευσης με πιθανότητες σε 17 (από 45) πεδία όλων των περιοχών. Η διαφορετικότητα στις προβλέψεις των συστημάτων το βασικού επιπέδου για ορισμένα πεδία είναι προφανώς μεγαλύτερη από μερικά άλλα, με αποτέλεσμα η συσσωρευμένη γενίκευση να αποδίδει καλύτερα.

Πίνακας 6.5 Κάθε κελί δείχνει τις στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, με βάση τη μετρική $F1$, σε όλα τα πεδία και των πέντε περιοχών, του συστήματος εξαγωγής πληροφορίας στη γραμμή, έναντι του αντιστοίχου της στήλης.

	Βασικό	MVotM	MVotF	PVotM	PVotF	Συς/ση απλή	Συσ/ση πιθαν.
Βασικό		14\6	13\9	13\9	10\18	0\11	1\17
MVotM	6\14		7\14	2\6	2\23	3\18	1\22
MVotF	9\13	14\7		15\7	8\16	3\12	1\20
PVotM	9\13	6\2	7\15		2\23	3\18	2\23
PVotF	18\10	23\2	16\8	23\2		7\9	2\13
Συσ/ση απλή	11\0	18\3	12\3	18\3	9\7		3\13
Συσ/ση πιθαν.	17\1	22\1	20\1	23\2	13\2	13\3	

Είναι αρκετά θετικό πάντως το γεγονός ότι δεν βλάπτεται από τη συσσωρευμένη γενίκευση η καλύτερη τιμή του βασικού επιπέδου, παρά μόνο σε ένα από τα 45 πεδία, όπου η μείωση στο $F1$ είναι αρκετά μικρή (από 98.16% σε 97.92%).

Επίσης, επιβεβαιώνεται και από τον Πίνακα 6.5 η ανωτερότητα της συσσωρευμένης γενίκευσης, έναντι της ψηφοφορίας, στα περισσότερα πεδία. Τα Παραρτήματα Α.4 έως Α.6 δείχνουν αναλυτικές τιμές και στις τρεις μετρικές αξιολόγησης για κάθε πεδίο από όλες τις τεχνικές ψηφοφορίας και συσσωρευμένης γενίκευσης. Η μέτρηση της διαφορετικότητας στις προβλέψεις των συστημάτων του βασικού επιπέδου, όπως φαίνεται στο Σχήμα 3.11 για κάθε περιοχή, μπορεί να γίνει και για κάθε πεδίο χωριστά. Η ανάλυση των δεδομένων σε μετα-επίπεδο, όπως φαίνεται στο Σχήμα 4.4 μπορεί επίσης να γίνει για κάθε πεδίο χωριστά, επιτρέποντας την εξήγηση των αποτελεσμάτων της εξαγωγής πληροφορίας σε ακόμα μεγαλύτερο βάθος.

6.1.3 Σύγκριση με βάση το υπολογιστικό κόστος

Εκτός από τα καλύτερα αποτελέσματα που επιτεύχθηκαν από τη συσσωρευμένη γενίκευση, το υπολογιστικό κόστος της τελευταίας είναι εξίσου ένα κρίσιμο ζήτημα όσον αφορά την πρακτική εφαρμογή σε προβλήματα εξαγωγής πληροφορίας. Το *κοινό κόστος εκπαίδευσης* των τεχνικών ψηφοφορίας και συσσωρευμένης γενίκευσης ισούται με το κόστος της εκπαίδευσης των συστημάτων του βασικού επιπέδου στα επισημειωμένα κείμενα, αθροιζόμενο με το κόστος μετατροπής των βαθμών εμπιστοσύνης σε πιθανότητες ορθότητας στις προβλέψεις των συστημάτων.

Από τα συστήματα που αξιολογήθηκαν σε βασικό επίπεδο στα πλαίσια αυτής της διατριβής, το σύστημα που βασίζεται στα HMMs είναι σημαντικά ταχύτερο κατά την εκπαίδευση από τα αντίστοιχα των αλγορίθμων BWI και $(LP)^2$, καθότι η εκπαίδευση βασίζεται σε απλοϊκούς υπολογισμούς, όπως εξηγείται στο Κεφάλαιο 3. Το σύστημα

που βασίζεται στον BWI είναι σημαντικά αργό, εξαιτίας της διαδικασίας ενδυνάμωσης κατά την εκπαίδευση, όπως εξηγείται με περισσότερες λεπτομέρειες στην εργασία [62]. Το σύστημα που βασίζεται στον $(LP)^2$ είναι επίσης αρκετά αργό, εξαιτίας κυρίως της εκμάθησης κανόνων διόρθωσης άλλων, που έχουν μαθευτεί σε προηγούμενο στάδιο, όπως περιγράφεται στην εργασία [26].

Η μετατροπή των βαθμών εμπιστοσύνης σε πιθανότητες ορθότητας στην έξοδο των συστημάτων απαιτεί, όπως περιγράφεται στην Ενότητα 5.6, την αξιολόγηση της εξόδου αυτής μέσω διασταυρωμένης επικύρωσης στα κείμενα εκπαίδευσης. Εφόσον η εκπαίδευση των συστημάτων που βασίζονται στους αλγορίθμους BWI και $(LP)^2$ είναι αργή, συμπεραίνουμε ότι και η διαδικασία αυτή είναι αργή, παρόλο που τα βήματα της διαδικασίας διασταυρωμένης επικύρωσης είναι τυπικά λιγότερα από τα βήματα της αντίστοιχης διαδικασίας για τη συσσωρευμένη γενίκευση. Μια εναλλακτική προσέγγιση για τη μείωση του υπολογιστικού κόστους κατά τη μετατροπή βαθμών εμπιστοσύνης σε πιθανότητες ορθότητας, είναι η αξιολόγηση της εξόδου των συστημάτων σε ένα ξεχωριστό τμήμα κειμένων, αντί της διενέργειας διασταυρωμένης επικύρωσης σε όλο το σύνολο των κειμένων εκπαίδευσης. Για παράδειγμα, το 80% των επισημειωμένων κειμένων θα μπορούσε να χρησιμοποιηθεί για την εκπαίδευση των συστημάτων του βασικού επιπέδου και το υπόλοιπο 20% για την αξιολόγηση και τη διαδικασία μετατροπής των βαθμών εμπιστοσύνης σε πιθανότητες ορθότητας.

Η σημαντικότερη διαφορά της ψηφοφορίας έναντι της συσσωρευμένης γενίκευσης, όσον αφορά το υπολογιστικό κόστος, αφορά την επιπλέον διαδικασία διασταυρωμένης επικύρωσης που απαιτεί η τελευταία στα επισημειωμένα κείμενα εκπαίδευσης, για την κατασκευή των διανυσμάτων χαρακτηριστικών και την εκπαίδευση ενός ταξινομητή σε μετα-επίπεδο. Η διαδικασία αυτή είναι αργή, όπως και η αντίστοιχη για τη μετατροπή βαθμών εμπιστοσύνης σε πιθανότητες ορθότητας. Τα κόστος εκπαίδευσης ενός ταξινομητή σε μετα-επίπεδο είναι σημαντικά μικρότερο από το αντίστοιχο ενός συστήματος του βασικού επιπέδου, παρά το γεγονός ότι η εκπαίδευση ορισμένων ταξινομητών (για παράδειγμα του *LogitBoost*) είναι περισσότερο χρονοβόρα από άλλους. Χαρακτηριστικό παράδειγμα είναι ότι η εκπαίδευση του συστήματος που βασίζεται στον BWI διαρκεί μερικές ώρες για τους φορητούς υπολογιστές, ενώ η εκπαίδευση ενός ταξινομητή σε μετα-επίπεδο διαρκεί λίγα λεπτά.

Όσον αφορά το κόστος κατά τη διαδικασία επαλήθευσης είναι σχεδόν κοινό τόσο για την ψηφοφορία όσο και τη συσσωρευμένη γενίκευση και ισούται με το κόστος εκτέλεσης των συστημάτων του βασικού επιπέδου, αθροιζόμενο με το κόστος συνδυασμού σε

μετα-επίπεδο. Το τελευταίο είναι σχεδόν αμελητέο για τις τεχνικές ψηφοφορίας, ενώ για τη συσσωρευμένη γενίκευση ισούται με το κόστος εκτέλεσης του εκπαιδευμένου ταξινομητή, το οποίο είναι και αυτό αρκετά μικρό.

Ο Πίνακας 6.6 δείχνει προσεγγιστικά το κόστος εκπαίδευσης για την ψηφοφορία και τη συσσωρευμένη γενίκευση. Το κόστος επαλήθευσης παραλείπεται καθώς είναι σχετικά μικρό και σχεδόν κοινό για τις δύο προσεγγίσεις. Τα πειράματα πραγματοποιήθηκαν σε μηχάνημα τύπου Pentium 4, 1.5 GHz, 523 MB Ram.

Πίνακας 6.6 Κόστος εκπαίδευσης (σε ώρες) για την ψηφοφορία και τη συσσωρευμένη γενίκευση σε κάθε περιοχή ενδιαφέροντος.

	Μαθήματα	Προγράμματα	Φορητοί	Αγγελίες	Σεμινάρια
Κόστος ψηφοφορίας	4	5	20	15	7
Κόστος συσσώρευσης	13	17	90	59	24

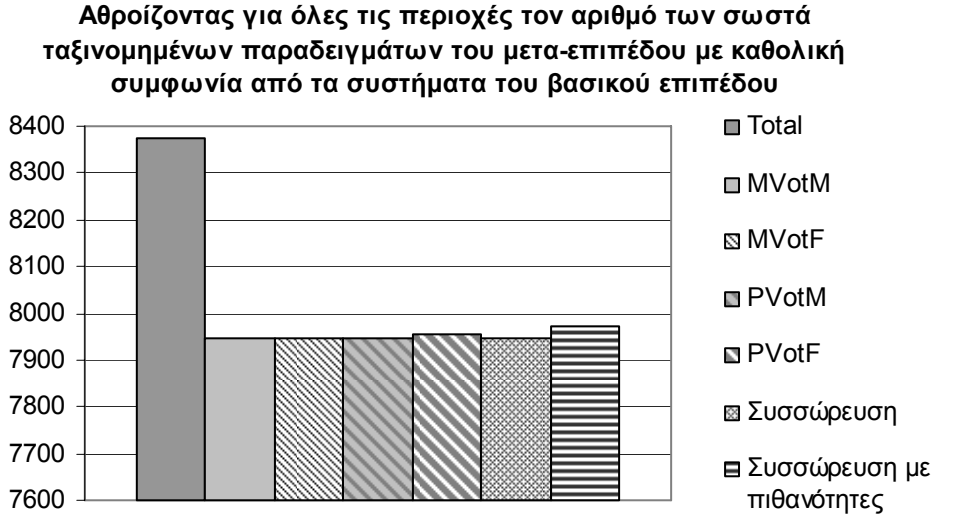
6.2 Ανάλυση αποτελεσμάτων με βάση τη διαφορετικότητα στις προβλέψεις των συστημάτων του βασικού επιπέδου

Στην Ενότητα 4.5 παρουσιάστηκε ένας διαχωρισμός των παραδειγμάτων του μετα-επιπέδου για κάθε θεματική περιοχή με βάση το βαθμό διαφωνίας ή συμφωνίας στις προβλέψεις των συστημάτων εξαγωγής πληροφορίας του βασικού επιπέδου. Στην ενότητα αυτή συγκρίνεται η ψηφοφορία με τη συσσωρευμένη γενίκευση με βάση την ανάλυση αυτή. Στόχος είναι μια περισσότερο λεπτομερής ανάλυση της συμπεριφοράς τόσο της ψηφοφορίας όσο και της συσσωρευμένης γενίκευσης, μελετώντας τα αποτελέσματα που επιτυγχάνουν ανάλογα με το πώς μεταβάλλεται ο βαθμός διαφωνίας/συμφωνίας στην έξοδο των συστημάτων του βασικού επιπέδου.

6.2.1 Ανάλυση με βάση την καθολική συμφωνία στις προβλέψεις των συστημάτων

Το Σχήμα 6.1 συγκρίνει όλες τις τεχνικές συνδυασμού συστημάτων εξαγωγής πληροφορίας, αθροιστικά και για τις πέντε θεματικές περιοχές ενδιαφέροντος, όταν όλα τα συστήματα του βασικού επιπέδου συμφωνούν στο ίδιο πεδίο για ένα τμήμα κειμένου. Οι περιπτώσεις αυτές των παραδειγμάτων του μετα-επιπέδου αντιστοιχούν στην αριστερή στήλη του Σχήματος 4.4 για κάθε θεματική περιοχή. Πρέπει να υπενθυμιστεί ότι κάθε παράδειγμα του μετα-επιπέδου μπορεί να ταξινομηθεί σε μια από τις τιμές $\{f^1 ... f^Q, false\}$, όπου $\{f^1 ... f^Q\}$ είναι τα σχετικά πεδία μιας θεματικής περιοχής και $false$

είναι μια ειδική τιμή που σημαίνει ότι το τμήμα κειμένου στο οποίο αντιστοιχεί το παράδειγμα δεν είναι σχετικό και δεν θα πρέπει επομένως να προβλεφθεί κάποιο πεδίο για αυτό από κανένα σύστημα του βασικού επιπέδου. Το Παράρτημα Β.1 δείχνει αναλυτικές τιμές για κάθε θεματική περιοχή χωριστά.



Σχήμα 6.1 Σύγκριση των τεχνικών συνδυασμού, όταν όλα τα συστήματα εξαγωγής πληροφορίας του βασικού επιπέδου συμφωνούν στο ίδιο πεδίο. Οι τιμές στον κάθετο άξονα είναι αθροισμένες για όλες τις θεματικές περιοχές και αντιστοιχούν στον αριθμό των παραδειγμάτων του μετα-επιπέδου.

Το Σχήμα 6.1 δείχνει ότι η συσσωρευμένη γενίκευση με χρήση πιθανοτήτων επιτυγχάνει ελαφρώς καλύτερα αποτελέσματα σε σχέση με τις υπόλοιπες μεθόδους συνδυασμού, ενώ το σχήμα ψηφοφορίας ακολουθεί *PVotF* σε απόδοση. Με άλλα λόγια, η συσσωρευμένη γενίκευση αποδεικνύεται ελαφρώς χρήσιμη ακόμα κι όταν οι προβλέψεις όλων των συστημάτων του βασικού επιπέδου ταυτίζονται. Εφόσον υπάρχει καθολική συμφωνία στις προβλέψεις των συστημάτων του βασικού επιπέδου, όλα τα σχήματα ψηφοφορίας, πλην του *PVotF*, επιστρέφουν το ίδιο πεδίο. Μόνο το *PVotF* διαθέτει την επιπρόσθετη επιλογή της απόρριψης μιας πρόβλεψης πεδίου, εάν η συνδυασμένη πιθανότητα για το πεδίο αυτό από όλα τα συστήματα είναι μικρότερη από συγκεκριμένο κατώφλι, κάτι που αποδείχτηκε ελαφρά πιο χρήσιμο για τα ερευνητικά προγράμματα και τους φορητούς υπολογιστές, σύμφωνα με το Παράρτημα Β.1.

Συγκρίνοντας το συνολικό αριθμό των παραδειγμάτων του μετα-επιπέδου όπου υπάρχει καθολική συμφωνία στις προβλέψεις από τα συστήματα του βασικού επιπέδου (πρώτη στήλη στο Σχήμα 6.1) και τον αριθμό των παραδειγμάτων εκείνων που έχουν ταξινομηθεί σωστά από όλες τις μεθόδους συνδυασμού, οδηγούμαστε στο ιδιαίτερα

ενδιαφέρον συμπέρασμα ένα πεδίο που προβλέπεται και από τα τρία συστήματα δεν είναι πάντα σωστό. Αυτό παρατηρήθηκε περισσότερο στις θεματικές περιοχές του παγκοσμίου ιστού, και οφείλεται αφενός στα σφάλματα κατά την επισημείωση των σελίδων (μειοψηφία των περιπτώσεων) και αφετέρου στο γεγονός ότι υπάρχουν χαρακτηριστικά του κειμένου που λανθασμένα αναγνωρίζουν και τα τρία συστήματα εξαγωγής πληροφορίας του βασικού επιπέδου (πλειοψηφία των περιπτώσεων). Για παράδειγμα, το “TFT” μπορεί να αντιστοιχεί σε ένα ξεχωριστό προϊόν οθόνης το οποίο περιγράφεται στην ίδια σελίδα με έναν φορητό ηλεκτρονικό υπολογιστή. Ομοίως για τα ερευνητικά προγράμματα, μερικοί καθηγητές ή μαθητές είναι πρώην μέλη ενός ερευνητικού προγράμματος, και επομένως δεν έχουν επισημειωθεί ως θετικά παραδείγματα από τον ειδικό της θεματικής περιοχής. Σε μερικές τέτοιες περιπτώσεις, η συσσωρευμένη γενίκευση με χρήση πιθανοτήτων μαθαίνει σωστά να ταξινομεί ως “false”, δηλαδή να απορρίπτει, τα αντίστοιχα διανύσματα χαρακτηριστικών. Βέβαια, δεν θα ήταν δυνατόν να αναμένουμε θεαματική βελτίωση δίχως την κωδικοποίηση περαιτέρω πληροφορίας στα διανύσματα χαρακτηριστικών του μετα-επιπέδου.

Οι παρατηρήσεις της προηγούμενης παραγράφου δείχνουν μια αδυναμία στα συστήματα εξαγωγής πληροφορίας του βασικού επιπέδου η οποία δεν είναι άγνωστη στη βιβλιογραφία. Η αδυναμία αυτή είναι ότι τα συστήματα αναλύουν κάθε κείμενο ως μια ακολουθία λεκτικών μονάδων και επομένως αγνοούν ιεραρχική πληροφορία που μπορεί να είναι διαθέσιμη σε μια HTML ή XML σελίδα του ιστού και που θα μπορούσε να αποβεί χρήσιμη για την εξαγωγή πληροφορίας. Η ιεραρχική αυτή πληροφορία μπορεί να είναι διαθέσιμη μέσω της μοντελοποίησης DOM *(Document Object Model, http://www.w3c.org)*. Επομένως, τα διαθέσιμα συστήματα εξαγωγής πληροφορίας εκμεταλλεύονται, κατά την εκπαίδευση και επαλήθευση, χαρακτηριστικά του κειμένου που αφορούν ακολουθίες λεκτικών μονάδων μέσα στα σχετικά (επισημειωμένα) τμήματα κειμένου καθώς και γύρω από αυτά, αποτυγχάνοντας να εκμεταλλευτούν ιεραρχική πληροφορία που είναι τυχόν διαθέσιμη. Για παράδειγμα, θα ήταν ενδιαφέρον να μπορούσαμε να εκμεταλλευτούμε ιεραρχική πληροφορία για να διαχωρίσουμε διαφορετικά προϊόντα φορητών υπολογιστών που περιγράφονται σε μια σελίδα.

Υπάρχουν εργασίες στη βιβλιογραφία οι οποίες εκμεταλλεύονται ιεραρχική πληροφορία για το πρόβλημα της εξαγωγής πληροφορίας, αλλά μειονεκτούν αφενός στην απαίτηση σε σημαντικό βαθμό χειρονακτικής εργασίας, και αφετέρου στη δυσκολία/έλλειψη προσαρμοστικότητας σε διαφορετικούς τύπους δομής κειμένου. Για παράδειγμα, ο αλγόριθμος STALKER [82] μαθαίνει κανόνες εξαγωγής πληροφορίας οι οποίοι μπορούν

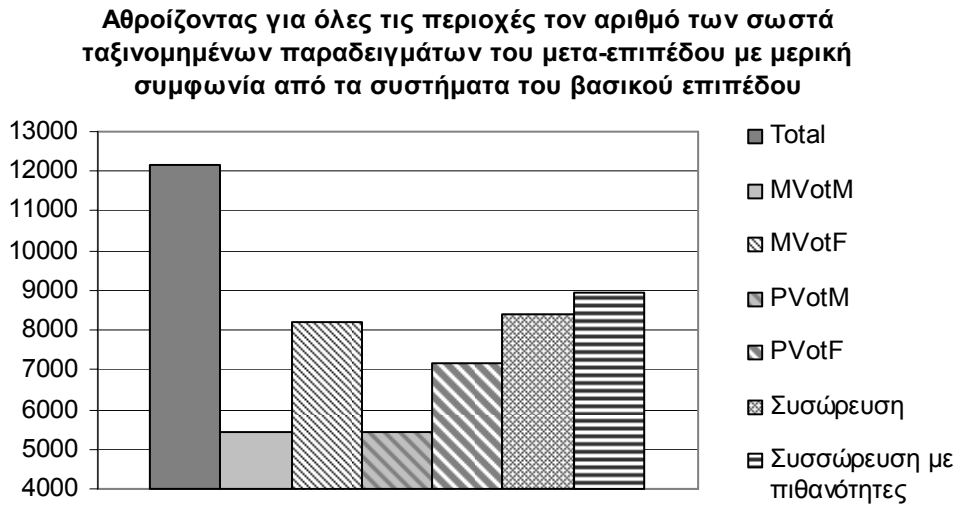
να αναγνωρίζουν συνεχόμενες ακολουθίες λεκτικών μονάδων, οι οποίες διαχωρίζονται από ενδιάμεσο κείμενο που δεν είναι σχετικό για την εξαγωγή. Κάτι τέτοιο, όμως, απαιτεί τη χειρονακτική κατασκευή ενός ειδικού φορμαλισμού, με διακριτικό τίτλο *Embedded Catalog (EC) formalism*, ο οποίος περιγράφει σελίδες του σιστού οι οποίες μοιράζονται την ίδια δομή στο περιεχόμενό τους. Διαφορετικοί τέτοιοι φορμαλισμοί θα πρέπει όμως να κατασκευαστούν χειρονακτικά για διαφορετικούς τύπους δομής.

Στην εργασία [34], ιεραρχική πληροφορία βασισμένη στη χρήση του DOM χρησιμοποιείται για εξαγωγή πληροφορίας από σελίδες του παγκοσμίου ιστού οι οποίες έχουν διαφορετική δομή, σε συνεργασία με μια οντολογία που κατασκευάζεται χειρονακτικά. Οι τεχνικές συνδυασμού συστημάτων που περιγράφονται και αξιολογούνται σε αυτή τη διατριβή θα μπορούσαν να συνεργαστούν με τις τεχνικές εξαγωγής πληροφορίας που περιγράφονται στην εργασία [34]. Για παράδειγμα, οι χειρονακτικά κατασκευασμένοι κανόνες για την αναγνώριση παραδειγμάτων σχετικών πεδίων τα οποία πρόκειται να εισαχθούν σε μια οντολογία, θα μπορούσαν να αντικατασταθούν από πιο πολύπλοκους και αποδοτικούς συνάμα κανόνες, οι οποίοι θα *μαθαίνονταν* από το συνδυασμό διαφορετικών συστημάτων εξαγωγής πληροφορίας. Κάτι τέτοιο θα ελάττωνε το κόστος κατασκευής/συντήρησης μιας οντολογίας. Αρχικές προσπάθειες χρήσης μηχανικής μάθησης σε εργασίες κατασκευής/συντήρησης οντολογιών υπάρχουν ήδη στη βιβλιογραφία [113].

6.2.2 Ανάλυση με βάση τη μερική συμφωνία στις προβλέψεις των συστημάτων

Το Σχήμα 6.2 συγκρίνει όλες τις τεχνικές συνδυασμού συστημάτων εξαγωγής πληροφορίας, αθροιστικά και για τις πέντε θεματικές περιοχές, όταν κάποια αλλά όχι όλα τα συστήματα του βασικού επιπέδου συμφωνούν στο ίδιο πεδίο για ένα τμήμα κειμένου, ενώ τα υπόλοιπα δεν προβλέπουν κάποιο πεδίο. Αφού αξιολογούνται τρία συστήματα, αυτό αντιστοιχεί σε περιπτώσεις που είτε τα δύο προβλέπουν το ίδιο πεδίο ενώ το τρίτο δεν προβλέπει τίποτα, είτε μόνο ένα σύστημα από τα τρία κάνει πρόβλεψη.

Το Σχήμα 6.2 δείχνει την ανωτερότητα της συσσωρευμένης γενίκευσης στην εκμετάλλευση της μερικής συμφωνίας στην έξοδο των συστημάτων του βασικού επιπέδου, η οποία μερική συμφωνία μπορεί να μεταφραστεί και ως διαφωνία. Για παράδειγμα “δύο συστήματα προβλέπουν, με διαφορετική πιθανότητα το καθένα, το πεδίο *ram* για το τμήμα κειμένου “256 MB”, ενώ το τρίτο σύστημα διαφωνεί και δεν προβλέπει κάποιο πεδίο”. Ο χειρισμός των αγνοούμενων τιμών από τη συσσωρευμένη γενίκευση δεν επηρεάζει σημαντικά τα αποτελέσματα, σύμφωνα και με την Ενότητα 5.7.



Σχήμα 6.2 Σύγκριση των τεχνικών συνδυασμού, όταν τα συστήματα του βασικού συμφωνούν μερικώς στο ίδιο πεδίο. Οι τιμές στον κάθετο άξονα είναι αθροισμένες για όλες τις περιοχές και αντιστοιχούν στον αριθμό των παραδειγμάτων σε μετα-επίπεδο.

Για την ανάλυση των αποτελεσμάτων του Σχήματος 6.2 σε μεγαλύτερο βάθος, το Σχήμα 6.3 συγκρίνει όλες τις μεθόδους συνδυασμού, ανάλογα με τον εάν μόνο ένα σύστημα από τα τρία του βασικού επιπέδου προβλέπει κάποιο πεδίο, ή ακριβώς δύο από τα τρία συστήματα συμφωνούν στο ίδιο πεδίο.

Σχήμα 6.3 Σύγκριση όλων των τεχνικών συνδυασμού συστημάτων εξαγωγής πληροφορίας, όταν μόνο ένα ή ακριβώς δύο από τα συστήματα συμφωνούν στο ίδιο πεδίο. Οι τιμές στον κάθετο άξονα είναι αθροισμένες για όλες τις θεματικές περιοχές και αντιστοιχούν στον αριθμό των παραδειγμάτων του μετα-επιπέδου.

Το Σχήμα 6.3 επιβεβαιώνει την υπεροχή της συσσωρευμένης γενίκευσης με χρήση πιθανοτήτων και στις δύο περιπτώσεις (μονή πρόβλεψη πεδίου ή δύο προβλέψεις στο

ίδιο πεδίο). Το σύνολο των στηλών στο αριστερό μέρος του σχήματος (μονή πρόβλεψη) επιβεβαιώνει τη συμπληρωματική συμπεριφορά των μεθόδων *MVotM/PVotM* με το *MVotF*. Πρέπει να υπενθυμίσουμε ότι τα σχήματα ψηφοφορίας *MVotM* και *PVotM* μοιράζονται σχεδόν το ίδιο μέγεθος στήλης αφού αγνοούμενες προβλέψεις παραβλέπονται ενώ πολύ αραιά τυχαίνουν διφορούμενες προβλέψεις πεδίων για ένα τμήμα κειμένου, όπως φαίνεται εξάλλου στο Σχήμα 4.4. Επομένως, στην περίπτωση μονής πρόβλεψης για ένα τμήμα κειμένου, το μέγεθος της στήλης των *MVotM/PVotM* ισούται με τον αριθμό των περιπτώσεων που το προβλεπόμενο πεδίο είναι σωστό.

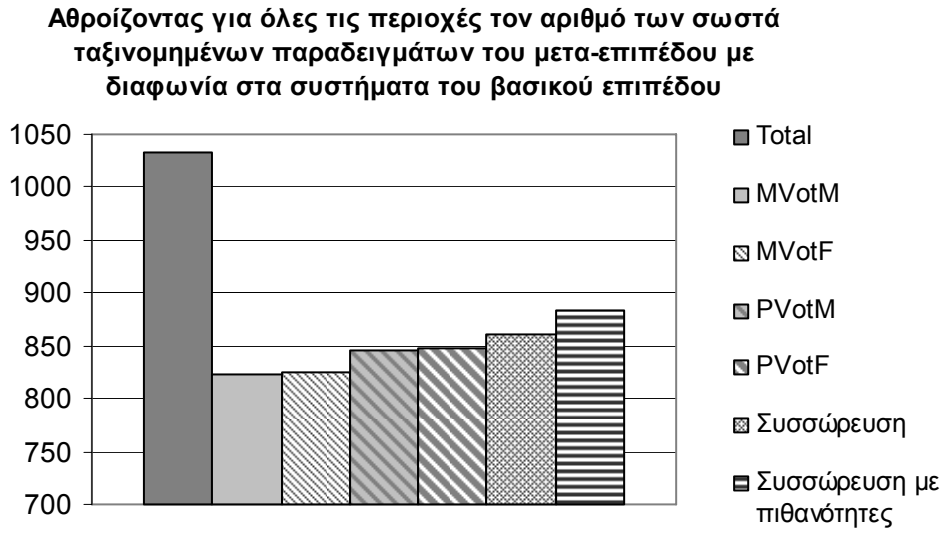
Το μέγεθος της στήλης για το *MVotF* ισούται με τον αριθμό των περιπτώσεων όπου το προβλεπόμενο πεδίο δεν είναι σωστό, όπου και επιστρέφεται η τιμή “false”. Το μεγάλο μέγεθος της στήλης για το *MVotF* στο αριστερό μέρος του Σχήματος 6.3, δείχνει ότι στην περίπτωση μονής πρόβλεψης πεδίου για ένα τμήμα κειμένου, δηλαδή πρόβλεψη μόνο του ενός από τα τρία διαθέσιμα συστήματα σε βασικό επίπεδο, η πρόβλεψη αυτή είναι περισσότερο πιθανό να είναι λανθασμένη.

Το Σχήμα 6.3 δείχνει επίσης ότι η συσσωρευμένη γενίκευση με χρήση πιθανοτήτων μαθαίνει κάτι περισσότερο από την συσσωρευμένη γενίκευση με χρήση ονομαστικών τιμών, καθώς και από ότι τα σχήματα *MVotM/PVotM* και *MVotF* προβλέπουν με συμπληρωματικό τρόπο, επιτυγχάνοντας μεγαλύτερη ακρίβεια στην ταξινόμηση των παραδειγμάτων του μετα-επιπέδου από ότι όλες οι υπόλοιπες τεχνικές συνδυασμού.

Όταν ακριβώς δύο συστήματα συμφωνούν στο ίδιο πεδίο ενώ το τρίτο δεν προβλέπει τίποτα, τότε όλα τα σχήματα ψηφοφορίας, πλην του *PVotF*, επιστρέφουν προφανώς το ίδιο πεδίο, όπως φαίνεται και από το δεξιό σύνολο στηλών του Σχήματος 6.3. Το *PVotF* τα πηγαίνει ελαφρώς καλύτερα, ενώ η συσσωρευμένη γενίκευση με χρήση ονομαστικών τιμών δεν έχει να επιδείξει κάτι πολύ πιο σημαντικό σε σύγκριση με την απόδοση των σχημάτων ψηφοφορίας. Από την άλλη πλευρά, η συσσωρευμένη γενίκευση με πιθανότητες μαθαίνει πάλι κάτι περισσότερο από την αντίστοιχη με ονομαστικές τιμές, και περισσότερο από ότι όλα τα σχήματα ψηφοφορίας προβλέπουν με προφανή τρόπο, επιτυγχάνοντας και πάλι μεγαλύτερη ακρίβεια στην ταξινόμηση των παραδειγμάτων του μετα-επιπέδου από ότι όλα τα υπόλοιπα σχήματα συνδυασμού.

6.2.3 Ανάλυση με βάση τη διαφωνία στις προβλέψεις των συστημάτων

Τέλος, το Σχήμα 6.4 συγκρίνει όλα τα σχήματα συνδυασμού και αθροιστικά για όλες τις θεματικές περιοχές, όταν υπάρχουν διφορούμενες προβλέψεις πεδίων, κάτι που αντιστοιχεί στη μεσαία στήλη για κάθε θεματική περιοχή στο Σχήμα 4.4.



Σχήμα 6.4 Σύγκριση όλων των τεχνικών συνδυασμού συστημάτων εξαγωγής πληροφορίας, όταν υπάρχει διαφωνία στις προβλέψεις πεδίων. Οι τιμές στον κάθετο άξονα είναι αθροισμένες για όλες τις θεματικές περιοχές και αντιστοιχούν στον αριθμό των παραδειγμάτων του μετα-επιπέδου.

Το Σχήμα 6.4 επιβεβαιώνει την ανωτερότητα της συσσωρευμένης γενίκευσης με πιθανότητες. Συνολικά τα Σχήματα 6.1 έως 6.4 δείχνουν ότι υπάρχει αρκετό περιθώριο βελτίωσης για περεταίρω βελτίωση των αποτελεσμάτων σε μετα-επίπεδο. Από την άλλη πλευρά, τα αποτελέσματα που παρουσιάζονται σε αυτή τη διατριβή για τις τεχνικές ψηφοφορίας και συσσωρευμένης γενίκευσης είναι αρκετά ενθαρρυντικά, λαμβάνοντας υπόψη την εκμετάλλευση απλής πληροφορίας σε μετα-επίπεδο, όπως οι ονομαστικές προβλέψεις των συστημάτων του βασικού επιπέδου, και οι πιθανότητες ορθότητας στις προβλέψεις αυτές. Από τη στιγμή που το πρόβλημα της εξαγωγής πληροφορίας έχει μετατραπεί σε ένα πρόβλημα ταξινόμησης σε μετα-επίπεδο, περαιτέρω πληροφορία μπορεί να κωδικοποιηθεί στα διανύσματα χαρακτηριστικών του μετα-επιπέδου.

6.2.4 Ανάλυση με βάση την ακρίβεια ταξινόμησης σε μετα-επίπεδο

Τα Σχήματα 6.1 έως 6.4 συγκρίνουν τις τεχνικές ψηφοφορίας και συσσωρευμένης γενίκευσης με βάση τον αριθμό των σωστά ταξινομημένων παραδειγμάτων του μετα-επιπέδου, ανάλογα με το διαφορετικό βαθμό ομοιότητας ή διαφωνίας στις προβλέψεις των συστημάτων του βασικού επιπέδου. Η *ακρίβεια ταξινόμησης σε μετα-επίπεδο*, μπορεί επομένως να οριστεί ως το ποσοστό των παραδειγμάτων του μετα-επιπέδου που έχουν σωστά ταξινομηθεί. Ο Πίνακας 6.7 συγκρίνει ξανά όλες τις τεχνικές συνδυασμού, με βάση τις στατιστικά σημαντικότερες *νίκες* έναντι *ηττών* στις πέντε θεματικές περιοχές ενδιαφέροντος, χρησιμοποιώντας την ακρίβεια ταξινόμησης.

Πίνακας 6.7 Στατιστικά σημαντικότερες *νίκες* έναντι *ηττών*, χρησιμοποιώντας την ακρίβεια ταξινόμησης, του συστήματος εξαγωγής πληροφορίας γραμμής, έναντι εκείνου της στήλης.

	MVotM	MVotF	PVotM	PVotF	Συσ/ση απλή	Συσ/ση πιθαν.
MVotM		0\4	0\1	0\5	0\5	0\5
MVotF	4\0		4\0	2\0	0\3	0\5
PVotM	1\0	0\4		0\5	0\5	0\5
PVotF	5\0	0\2	5\0		0\3	0\5
Συσ/ση απλή	5\0	3\0	5\0	3\0		0\4
Συσ/ση πιθαν.	5\0	5\0	5\0	5\0	4\0	

Ο Πίνακας 6.7 δείχνει την υπεροχή της συσσωρευμένης γενίκευσης σε όλες τις θεματικές περιοχές, σε σύγκριση με τις υπόλοιπες τεχνικές συνδυασμού. Αυτό όμως έρχεται σε αντίθεση με τα αποτελέσματα του Πίνακα 6.4 που δείχνουν ότι το σχήμα ψηφοφορίας *PVotF* επιτυγχάνει συγκρίσιμα αποτελέσματα με τη συσσωρευμένη γενίκευση σε δύο από τις πέντε περιοχές, και συγκεκριμένα στα μαθήματα της επιστήμης υπολογιστών και στους φορητούς ηλεκτρονικούς υπολογιστές. Επιπλέον, ο Πίνακας 6.7 δείχνει ότι το *MVotF* είναι το καλύτερο σχήμα ψηφοφορίας, σε αντίθεση με τον Πίνακα 6.4 που δείχνει ότι το *PVotF* είναι η καλύτερη τεχνική ψηφοφορίας στις περισσότερες θεματικές περιοχές ενδιαφέροντος.

Τα παραπάνω αντικρουόμενα συμπεράσματα οφείλονται στις διαφορετικές μετρικές αξιολόγησης που χρησιμοποιήθηκαν στους δύο πίνακες. Η μέτρηση των στατιστικά σημαντικότερων *νικών/ ηττών* στον Πίνακα 6.4 γίνεται με χρήση της *μικρο-υπολογιστικής (micro-average)* $F1$, μετρημένης σε όλα τα σχετικά πεδία $\{f^1...f^Q\}$ για κάθε περιοχή. Από την άλλη πλευρά, η ακρίβεια ταξινόμησης σε μετα-επίπεδο που δείχνει ο Πίνακας 6.7 ορίζεται σε όλες τις δυνατές τιμές $\{f^1...f^Q, false\}$ που μπορεί να ταξινομηθεί ένα παράδειγμα του μετα-επιπέδου, όπου συμπεριλαμβάνεται και η τιμή “false”. Στην πραγματικότητα, η ακρίβεια ταξινόμησης σε μετα-επίπεδο ταυτίζεται με την μικρο-υπολογιστική ακρίβεια υπολογισμένη σε όλες τις τιμές $\{f^1...f^Q, false\}$.

Η ακρίβεια ταξινόμησης είναι μια τυπική μετρική που χρησιμοποιείται για τη σύγκριση πολλών ταξινομητών και υπολογίζεται με βάση όλες τις σχετικές κλάσεις μιας θεματικής περιοχής. Στην εξαγωγή πληροφορίας, όμως, η τιμή κλάσης “false” έχει μια ιδιαίτερη σημασιολογία, αφού καμία τέτοια τιμή δεν επισημειώνεται στα κείμενο από τον ειδικό της θεματικής περιοχής. Ομοίως, κανένα σύστημα σε βασικό επίπεδο δεν προβλέπει την τιμή “false” για κάποιο τμήμα κειμένου. Η τιμή “false”, όμως, είναι μια δυνατή επιλογή για τον ταξινομητή σε μετα-επίπεδο, δείχνοντας με την επιλογή αυτή ότι κανένα από τα συστήματα του βασικού επιπέδου δεν έχει κάνει σωστή πρόβλεψη.

Η αξιολόγηση στην εξαγωγή πληροφορίας λαμβάνει χώρα συγκρίνοντας το χειρονακτικά συμπληρωμένο σχεδιότυπο για μια σελίδα κειμένου με το αντίστοιχο που συμπληρώνεται από ένα σύστημα σε βασικό ή σε μετα-επίπεδο. Επομένως, οι μετρικές αξιολόγησης *ανάκληση*, *ακρίβεια* και $F1$ υπολογίζονται με βάση μόνο τα σχετικά πεδία $\{f^1 ... f^Q\}$ της θεματικής περιοχής, αγνοώντας έτσι την τιμή “false”.

6.2.5 Ανάλυση με βάση την απουσία πρόβλεψης από όλα τα συστήματα

Υπενθυμίζεται ξανά ότι κάθε παράδειγμα του μετα-επιπέδου αντιστοιχεί σε ένα τμήμα κειμένου που έχει αναγνωριστεί ως σχετικό από τουλάχιστον ένα σύστημα του βασικού επιπέδου. Ένα όμως ένα σχετικό τμήμα κειμένου δεν έχει αναγνωριστεί από κανένα σύστημα, τότε φυσικά δεν υπάρχει καμία πιθανότητα να αναγνωριστεί σε μετα-επίπεδο, είτε μέσω ψηφοφορίας είτε μέσω συσσωρευμένης γενίκευσης. Στην περίπτωση της τελευταίας που απαιτεί μια εσωτερική διαδικασία διασταυρωμένης επικύρωσης στο σύνολο των επισημειωμένων κειμένων εκπαίδευσης του βασικού επιπέδου, ώστε να κατασκευαστεί το σύνολο των διανυσμάτων χαρακτηριστικών του μετα-επιπέδου, κάτι τέτοιο οδηγεί σε απώλεια πληροφορίας για τον ταξινομητή που θα εκπαιδευτεί στα νέα διανύσματα. Για παράδειγμα, εάν το τμήμα κειμένου “256 MB” δεν έχει αναγνωριστεί από κανένα σύστημα του βασικού επιπέδου, τότε δεν θα κατασκευαστεί προφανώς κάποιο διάνυσμα χαρακτηριστικών για το συγκεκριμένο τμήμα.

Ο Πίνακας 6.8 δείχνει το ποσοστό των σχετικών τμημάτων κειμένου, με βάση τα χειρονακτικά συμπληρωμένα σχεδιότυπα, που έχουν αναγνωριστεί από τουλάχιστον ένα σύστημα εξαγωγής πληροφορίας του βασικού επιπέδου, τόσο κατά την εκπαίδευση (δηλαδή κατά την εσωτερική διαδικασία διασταυρωμένης επικύρωσης στα κείμενα εκπαίδευσης) όσο και κατά τη διαδικασία επαλήθευσης.

Πίνακας 6.8 Ποσοστό (%) των σχετικών (επισημειωμένων) τμημάτων κειμένου τα οποία έχουν αναγνωριστεί από τουλάχιστον ένα σύστημα του βασικού επιπέδου.

	Μαθήματα	Προγράμματα	Φορητοί	Αγγελίες	Σεμινάρια
Εκπαίδευση	63.26	85.34	75.11	91.88	86.26
Επαλήθευση	81.44	85.88	77.06	93.02	88.66

Στη συσσωρευμένη γενίκευση για κοινά προβλήματα ταξινόμησης δεν υπάρχει χάσιμο πληροφορίας, αφού υπάρχει ένα προς ένα αντιστοιχία μεταξύ των διανυσμάτων χαρακτηριστικών του βασικού και του μετα-επιπέδου και επομένως οι τιμές των χαρακτηριστικών κλάσης είναι ταυτόσημες σε βασικό και σε μετα-επίπεδο.

Οι τιμές του Πίνακα 6.8 για τη διαδικασία επαλήθευσης προσδιορίζουν μια μέγιστη τιμή για την ανάκληση που μπορεί να επιτευχθεί σε μετα-επίπεδο, είτε μέσω ψηφοφορίας

είτε μέσω συσσωρευμένης γενίκευσης, αφού η αξιολόγηση γίνεται συγκρίνοντας τα χειρονακτικά συμπληρωμένα σχεδιότυπα με τα αντίστοιχα που έχουν συμπληρωθεί σε μετα-επίπεδο. Η ιδανική περίπτωση θα ήταν όλα τα χειρονακτικά επισημειωμένα τμήματα κειμένου να έχουν αναγνωριστεί από ένα τουλάχιστον σύστημα του βασικού επιπέδου. Κάτι τέτοιο θα εξασφάλιζε τη δημιουργία σε μετα-επίπεδο ενός διανύσματος χαρακτηριστικών για κάθε σχετικό τμήμα, και άρα μηδενική απώλεια πληροφορίας.

Παρά την προφανή απώλεια πληροφορίας στην εξαγωγή πληροφορία οι τεχνικές ψηφοφορίας και συσσωρευμένης γενίκευσης καταφέρνουν και βελτιώνουν σε μετα-επίπεδο την απόδοση των συστημάτων του βασικού επιπέδου. Συγκρίνοντας, τέλος, τις καλύτερες τιμές για την ανάκληση από τις τεχνικές ψηφοφορίας και συσσωρευμένης γενίκευσης (Πίνακας 5.8) και των τιμών του Πίνακα 6.8, συμπεραίνουμε ότι υπάρχει χώρος για περαιτέρω βελτίωση της ανάκλησης σε μετα-επίπεδο κι επομένως της συνολικής απόδοσης στην εξαγωγή πληροφορίας και στις πέντε θεματικές περιοχές.

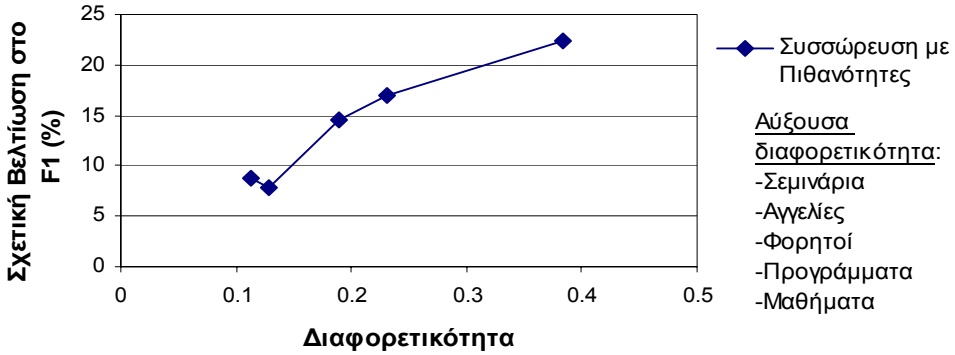
6.2.6 Σχετική βελτίωση σε μετα-επίπεδο, ανάλογα με τη διαφορετικότητα

Στην Ενότητα 3.5.2 μετρήθηκε η ομοιότητα στις προβλέψεις κάθε ζεύγους (E^i, E^j) συστημάτων του βασικού επιπέδου, ως η συνδυασμένη πιθανότητα το σύστημα E^i να κάνει σωστή πρόβλεψη, δοθέντος του ότι το σύστημα E^j προβλέπει επίσης σωστά. Η μέση ομοιότητα, επομένως στις προβλέψεις και των τριών συστημάτων του βασικού επιπέδου ορίζεται ως ο μέσος όρος της ομοιότητας όλων των ζευγαριών (E^i, E^j). Μπορούμε λοιπόν να ορίσουμε εύκολα τη μετρική *μέση διαφορετικότητα (average diversity)* των συστημάτων του βασικού επιπέδου, σύμφωνα με την εξίσωση 6.1.

$$\text{Μέση διαφορετικότητα} = \frac{1}{|E|(|E|-1)} \sum_{i \neq j} P(E^i = wrong \mid E^j = correct) \qquad (6.1)$$

όπου $|E|$ ο συνολικός αριθμός συστημάτων του βασικού επιπέδου.

Το Σχήμα 6.5 δείχνει τη *σχετική βελτίωση (relative improvement)* στο $F1$ (%) που επιτυγχάνει η συσσωρευμένη γενίκευση με πιθανότητες, σε σχέση με όλα τα συστήματα σε βασικό επίπεδο (δηλαδή ο μέσος όρος) και για κάθε περιοχή, σε σχέση με τη μετρική της 6.1. Το Σχήμα δείχνει τη σαφή τάση για βελτίωση στην απόδοση σε μετα-επίπεδο, όσο αυξάνεται η μέση διαφορετικότητα στις προβλέψεις των συστημάτων σε βασικό επίπεδο. Μοναδική μικρή παραφωνία αποτελεί η περιοχή των αγγελιών, όπου οφείλεται εν μέρει στο ότι είναι η μοναδική περιοχή όπου η χρήση πιθανοτήτων δεν αποδείχτηκε ιδιαίτερα ωφέλιμη κατά το χρήση συσσωρευμένης γενίκευσης.



Σχήμα 6.5 Σχέση μεταξύ (μέσης) σχετικής βελτίωση (%) στο $F1$, από τη συσσωρευμένη γενίκευση με πιθανότητες σε σχέση με το βασικό επίπεδο και μέσης διαφορετικότητας.

6.3 Συμπεράσματα

Η συσσωρευμένη γενίκευση με πιθανότητες αποδείχτηκε ανώτερη της ψηφοφορίας με πιθανότητες, που θέτει ένα κατώφλι στην αποδοχή προβλέψεων *(PVotF)*, στις τρεις από τις πέντε θεματικές περιοχές. Στις άλλες δύο, η συσσωρευμένη γενίκευση επιτυγχάνει πάλι ανώτερο $F1$ από το *PVotF*, που αξιολογήθηκε ως η καλύτερη τεχνική ψηφοφορίας, αλλά οι διαφορές μετρήθηκαν ως στατιστικά μη σημαντικές. Η συσσωρευμένη γενίκευση με πιθανότητες επιτυγχάνει μεγαλύτερη *ακρίβεια* από το *PVotF* και στις πέντε περιοχές, ενώ το *PVotF* υπερτερεί στην *ανάκληση*. Σκεπτόμενοι το επιπλέον υπολογιστικό κόστος της συσσωρευμένης γενίκευσης, το *PVotF* μπορεί να θεωρηθεί ως μια αρκετά ικανοποιητική τεχνική συνδυασμού, αφού βελτιώνει τα αποτελέσματα του βασικού επιπέδου στις περισσότερες περιοχές. Η συσσωρευμένη γενίκευση, όμως, βελτιώνει την απόδοση των καλύτερων συστημάτων του βασικού επιπέδου σε όλες τις περιοχές, ακόμα και σε εκείνες (αγγελίες και σεμινάρια) όπου τα περιθώρια βελτίωσης σε μετα-επίπεδο είναι περιορισμένα.

Η ανάλυση των αποτελεσμάτων με βάση τη διαφορετικότητα στην έξοδο των συστημάτων του βασικού επιπέδου, έδειξε ότι σε περιπτώσεις καθολικής συμφωνίας στις προβλέψεις των συστημάτων, η συσσωρευμένη γενίκευση με πιθανότητες οδηγεί σε ελαφρώς καλύτερα αποτελέσματα εξαγωγής σε μετα-επίπεδο. Αυτό σημαίνει ότι μπορεί να αποφανθεί σωστά σε κάποιες περιπτώσεις όπου όλα τα συστήματα του βασικού επιπέδου κάνουν λανθασμένες προβλέψεις. Επίσης σε περιπτώσεις όπου είτε μόνο ένα από τα τρία διαθέσιμα συστήματα σε βασικό επίπεδο κάνει πρόβλεψη, είτε ακριβώς δύο συστήματα συμφωνούν στο ίδιο πεδίο, η συσσωρευμένη γενίκευση με πιθανότητες επιτυγχάνει τα καλύτερα αποτελέσματα εξαγωγής από όλες τις τεχνικές συνδυασμού. Στις υπόλοιπες, λιγότερες, περιπτώσεις διαφωνίας στις προβλέψεις πεδίων, η συσσωρευμένη γενίκευση με πιθανότητες αναδεικνύεται και πάλι νικήτρια.

ΚΕΦΑΛΑΙΟ 7

ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

Οι τεχνικές ψηφοφορίας και συσσωρευμένης γενίκευσης ή συσσώρευσης, έχουν αποδειχτεί ιδιαίτερα αποτελεσματικές για το συνδυασμό πολλαπλών αλγορίθμων μηχανικής μάθησης. Παρόλα αυτά, η εφαρμογή τους έχει περιοριστεί σχεδόν αποκλειστικά σε κοινά προβλήματα ταξινόμησης. Η διατριβή αυτή ερεύνησε την εφαρμογή των τεχνικών αυτών σε προβλήματα εξαγωγής πληροφορίας στον ιστό αλλά και στον ευρύτερο χώρο του διαδικτύου, αναδεικνύοντας παράλληλα και την αποτελεσματικότητά τους σε πλήθος περιοχών και χρησιμοποιώντας γνωστούς αλγορίθμους σε βασικό και σε μετα-επίπεδο.

Για το συνδυασμό διαφορετικών συστημάτων εξαγωγής πληροφορίας προτάθηκε μια νέα μεθοδολογία, η οποία μετατρέπει το πρόβλημα της εξαγωγής πληροφορίας σε ένα πρόβλημα κοινής ταξινόμησης σε μετα-επίπεδο κι επιτρέπει την εφαρμογή τεχνικών ψηφοφορίας και συσσωρευμένης γενίκευσης. Η νέα μεθοδολογία επιτρέπει επίσης το συνδυασμό πληθώρας συστημάτων σε βασικό επίπεδο, καθώς δεν εξαρτάται από το πώς κάθε σύστημα μοντελοποιεί το πρόβλημα της εξαγωγής πληροφορίας. Τεχνικές ψηφοφορίας και συσσωρευμένης γενίκευσης ορίστηκαν και αξιολογήθηκαν στα πλαίσια της νέας μεθοδολογίας, αναδεικνύοντας ταυτόχρονα την αποτελεσματικότητά τους.

Η Ενότητα 7.1 συνοψίζει τα βασικά συμπεράσματα της παρούσας διατριβής. Στην Ενότητα 7.2 συζητούνται βελτιώσεις και επεκτάσεις των μεθόδων συνδυασμού συστημάτων εξαγωγής πληροφορίας που προτάθηκαν σε αυτή τη διατριβή.

7.1 Συμπεράσματα

Η αξιολόγηση ανέδειξε ότι τόσο η χρήση ψηφοφορίας όσο και η χρήση συσσωρευμένης γενίκευσης επιτυγχάνουν καλύτερα αποτελέσματα όταν βασίζονται στις πιθανότητες ορθότητας των συστημάτων του βασικού επιπέδου. Επίσης, τόσο η ψηφοφορία όσο και η συσσωρευμένη γενίκευση εκμεταλλεύονται επιτυχώς τη διαφορετικότητα στις προβλέψεις των συστημάτων εξαγωγής πληροφορίας του βασικού επιπέδου, οδηγώντας σε καλύτερα αποτελέσματα στην εξαγωγή πληροφορίας σε μετα-επίπεδο.

Η διενέργεια ψηφοφορίας με χρήση πιθανοτήτων, θέτοντας ένα όριο στην τελική πιθανότητα για την αποδοχή ή όχι μιας πρόβλεψης, αναδείχτηκε ως μια ιδιαίτερα αποτελεσματική μέθοδος συνδυασμού συστημάτων στις περισσότερες θεματικές

περιοχές, οδηγώντας σε καλύτερα αποτελέσματα από τα συστήματα του βασικού επιπέδου. Η συσσωρευμένη γενίκευση με χρήση πιθανοτήτων, από την άλλη πλευρά, αναδείχτηκε ως μια καθολικά αποτελεσματική μέθοδος συνδυασμού συστημάτων εξαγωγής πληροφορίας σε όλες τις περιοχές που εξετάστηκαν, οδηγώντας πάντα σε καλύτερα αποτελέσματα από τα συστήματα του βασικού επιπέδου.

Η συσσωρευμένη γενίκευση αποδείχτηκε ανώτερη της ψηφοφορίας και στις πέντε θεματικές περιοχές. Μόνο σε δύο περιοχές οι διαφορές μετρήθηκαν ως στατιστικά μη σημαντικές. Η συσσωρευμένη γενίκευση βελτιώνει πάντα την ακρίβεια σε μετα-επίπεδο, σε σχέση με τα καλύτερο σύστημα εξαγωγής του βασικού επιπέδου. Η ανάκληση επίσης βελτιώνεται στις περισσότερες περιοχές. Στις περιοχές όπου η ψηφοφορία επιτυγχάνει συγκρίσιμα αποτελέσματα με τη συσσωρευμένη γενίκευση, η τελευταία επιτυγχάνει πιο ακριβείς προβλέψεις. Αναλογίζοντας το επιπλέον υπολογιστικό κόστος της συσσωρευμένης γενίκευσης, η διενέργεια ψηφοφορίας μπορεί να θεωρηθεί μια αρκετά καλή λύση. Όμως το τελικό συμπέρασμα είναι ότι η συσσωρευμένη γενίκευση είναι η καλύτερη τεχνική συνδυασμού από όλες όσες αξιολογήθηκαν σε αυτή τη διατριβή, συμπεριλαμβανομένης και της πολυστρατηγικής μάθησης [48], της μοναδικής μέχρι τώρα τεχνικής συνδυασμού συστημάτων εξαγωγής στη διεθνή βιβλιογραφία.

Ιδιαίτερο ενδιαφέρον παρουσιάζουν κάποιες από τις τεχνικές ψηφοφορίας που ορίστηκαν σε αυτή τη διατριβή, παρόλο που δεν οδηγούν στα καλύτερα αποτελέσματα εξαγωγής σε μετα-επίπεδο. Η συσσωρευμένη γενίκευση με χρήση πιθανοτήτων είναι η καλύτερη μέθοδος συνδυασμού σύμφωνα με τη μετρική $F1$. Από την άλλη πλευρά, οι τεχνικές ψηφοφορίας με χρήση ονομαστικών και πιθανοτικών τιμών που παραβλέπουν τυχόν αγνοούμενες τιμές από τα συστήματα του βασικού επιπέδου επιτυγχάνουν τη μεγαλύτερη *ανάκληση* σε μετα-επίπεδο. Παρόλο που στην εξαγωγή πληροφορίας μας ενδιαφέρουν περισσότερο οι μετρικές της ακρίβειας και του $F1$, υπάρχουν εντούτοις εφαρμογές όπου η επίτευξη μεγαλύτερης ανάκλησης είναι περισσότερο χρήσιμη. Στην εργασία [113] για παράδειγμα, παρουσιάζεται μια μεθοδολογία *εμπλουτισμού οντολογίας με παραδείγματα (ontology population)*, κατά την οποία ένα σύστημα εξαγωγής πληροφορίας αναγνωρίζει παραδείγματα πεδίων τα οποία δεν υπάρχουν σε μια οντολογία για μια θεματική περιοχή. Ένας ειδικός αναγνωρίζει τα παραδείγματα αυτά τα οποία και εμπλουτίζουν την οντολογία, απομακρύνοντας ταυτόχρονα τα λανθασμένα παραδείγματα. Σε μια τέτοια μεθοδολογία, είναι συμφέρουσα η χρήση μεθόδων συνδυασμού που οδηγούν σε μεγαλύτερη ανάκληση σε μετα-επίπεδο, άρα και

σε καλύτερο εμπλουτισμό της οντολογίας, αφού υπάρχει ο ειδικός που βελτιστοποιεί την ακρίβεια, απομακρύνοντας τα λανθασμένα παραδείγματα.

Τα αποτελέσματα που επιτεύχθηκαν από την ψηφοφορία και τη συσσωρευμένη γενίκευση είναι ιδιαίτερα θετικά, λαμβάνοντας υπόψη την απλοϊκότητα των διανυσμάτων χαρακτηριστικών του μετα-επιπέδου. Απλές ονομαστικές προβλέψεις και πιθανοτικές εκτιμήσεις ορθότητας των προβλέψεων αυτών ήταν τα μόνα δεδομένα που χρησιμοποιήθηκαν στην αναπαράσταση των διανυσμάτων του μετα-επιπέδου.

Μια σημαντική συνεισφορά της διατριβής αυτής, υπήρξε επίσης η ανάλυση των αποτελεσμάτων των τεχνικών ψηφοφορίας και συσσωρευμένης γενίκευσης, με βάση τη διαφορετικότητα στις προβλέψεις των συστημάτων του βασικού επιπέδου. Στόχος ήταν να διερευνηθούν οι διάφορες πτυχές επιτυχίας των τεχνικών συνδυασμού. Το γεγονός ότι οι τεχνικές συνδυασμού εκμεταλλεύτηκαν τη διαφορετικότητα στις προβλέψεις των συστημάτων του βασικού επιπέδου, είναι αναμενόμενο. Ενδιαφέρον αποτελεί όμως το γεγονός ότι η διαφορετικότητα αυτή αποτελείται στη συντριπτική πλειοψηφία της από περιπτώσεις όπου κάποια (αλλά όχι όλα) συστήματα προβλέπουν το ίδιο πεδίο ενώ τα υπόλοιπα δεν κάνουν πρόβλεψη. Περιπτώσεις διαφωνίας στα προβλεπόμενα πεδία δεν είναι συχνές αλλά και σε αυτές υπερέχει η συσσωρευμένη γενίκευση. Ιδιαίτερα ενδιαφέρον αποτελεί επίσης η ελαφριά ανωτερότητα της συσσωρευμένης γενίκευσης ακόμα και όταν τα συστήματα του βασικού επιπέδου συμφωνούν στις προβλέψεις τους.

Η ανάλυση των αποτελεσμάτων έδειξε επίσης ότι για το συνδυασμό μέσω συσσωρευμένης γενίκευσης, είναι προτιμότερη η επιλογή συστημάτων εξαγωγής πληροφορίας με μεγαλύτερη *ανάκληση*, καθότι η *ακρίβεια* βελτιώνεται πάντα σε μετα-επίπεδο. Μεγαλύτερη ανάκληση σημαίνει μεγαλύτερη πιθανότητα για αύξηση της διαφορετικότητας μεταξύ των συστημάτων του βασικού επιπέδου, προς όφελος της συσσωρευμένης γενίκευσης. Μεγαλύτερη ανάκληση επίσης σημαίνει μεγαλύτερη πιθανότητα για μείωση των περιπτώσεων όπου σχετικά παραδείγματα πεδίων δεν έχουν αναγνωριστεί από κανένα σύστημα του βασικού επιπέδου, άρα και μείωση της απώλειας πληροφορίας σε μετα-επίπεδο για τη συσσωρευμένη γενίκευση.

Η χρήση προσαρμοστικών συστημάτων εξαγωγής είναι μια σημαντική προοπτική για τη συμπλήρωση του στόχου του σημασιολογικού ιστού, δηλαδή τη δημιουργία περιεχομένου κατανοητού από τις μηχανές, καθώς έχουν εφαρμογή τόσο σε ισχυρά δομημένο όσο και σε λιγότερο δομημένο κείμενο του ιστού. Τα συμπεράσματα της παρούσας διατριβής συνεισφέρουν στην κατεύθυνση αυτή καθώς αναδεικνύουν την αποτελεσματικότητα του συνδυασμού διαφορετικών προσαρμοστικών συστημάτων.

7.2 Μελλοντική εργασία

Από τη στιγμή που το πρόβλημα της εξαγωγής πληροφορίας μετατράπηκε σε ένα κοινό πρόβλημα ταξινόμησης σε μετα-επίπεδο, υπάρχει πληθώρα ευκαιριών για την περαιτέρω βελτίωση των αποτελεσμάτων. Εκτός από τις ονομαστικές τιμές και τις πιθανότητες ορθότητας στις προβλέψεις πεδίων, περαιτέρω πληροφορία μπορεί να κωδικοποιηθεί στα διανύσματα χαρακτηριστικών του μετα-επιπέδου για να δικαιολογηθεί το επιπλέον υπολογιστικό κόστος της συσσωρευμένης γενίκευσης σε σχέση με την ψηφοφορία. Για παράδειγμα, στους φορητούς υπολογιστές, παραδείγματα του πεδίου “ταχύτητα επεξεργαστή” τυπικά εμφανίζονται αμέσως μετά τα αντίστοιχα του πεδίου “όνομα επεξεργαστή”, ενώ παραδείγματα του πεδίου “ram” τυπικά ακολουθούν.

Εκμεταλλευόμενοι τέτοιου είδους εξαρτήσεις μεταξύ των πεδίων, ή άλλων πιθανών πηγών πληροφορίας, μπορούμε να οδηγηθούμε σε χρήσιμα χαρακτηριστικά για τα διανύσματα του μετα-επιπέδου, προς όφελος των ταξινομητών. Επίσης, η *ανομοιογένεια των τιμών κλάσης* (*class imbalance*, [59]) στα διανύσματα χαρακτηριστικών του μετα-επιπέδου είναι ένα πρόβλημα μπορεί να διερευνηθεί, αφού ορισμένα πεδία έχουν επισημειωμένα πολύ περισσότερα παραδείγματα μέσα στα κείμενα από άλλα. Ο συνδυασμός ταξινομητών σε δεύτερο μετα-επίπεδο, είτε μέσω ψηφοφορίας είτε μέσω συσσωρευμένης γενίκευσης, μπορεί επίσης να διερευνηθεί για την περαιτέρω βελτίωση των αποτελεσμάτων στην εξαγωγή πληροφορίας.

Μια εναλλακτική στρατηγική συνδυασμού μπορεί να ακολουθηθεί, κατά την οποία κάθε σχετικό πεδίο μιας περιοχής αντιμετωπίζεται χωριστά κατά το συνδυασμό [48]. Στην περίπτωση αυτή, μια ξεχωριστή διαδικασία διασταυρωμένης επικύρωσης θα λάμβανε χώρα στα δεδομένα εκπαίδευσης για κάθε πεδίο χωριστά. Το πρόβλημα θα μετασχηματιζόταν έτσι σε μια διαδικασία δυαδικής μάθησης σε μετα-επίπεδο, όπου κάθε παράδειγμα σε μετα-επίπεδο θα έπρεπε είτε να ταξινομηθεί ως σχετικό του συγκεκριμένου πεδίου ή να απορριφθεί. Μια τέτοια στρατηγική συνδυασμού θα αντιμετώπιζε ευκολότερα έναν γενικότερο περιορισμό της διασταυρωμένης επικύρωσης σε αρχεία κειμένου, αυτόν της *διαστρωμάτωσης (stratification)*. Σε προβλήματα ταξινόμησης, διατηρείται μια παρόμοια κατανομή των κλάσεων στα διανύσματα κάθε βήματος της διασταυρωμένης επικύρωσης. Στην εξαγωγή πληροφορίας, όμως, δε γίνεται να ισχύει κάτι τέτοιο, αφού χειριζόμαστε επισημειωμένα κείμενα και όχι διανύσματα χαρακτηριστικών. Επομένως, σε κάθε κείμενο υπάρχει τυπικά μια διαφορετική κατανομή των παραδειγμάτων πεδίων, καθιστώντας ιδιαίτερα δύσκολη τη διατήρηση παρόμοιας κατανομής στα κείμενα σε κάθε βήμα της διασταυρωμένης

επικύρωσης. Από την άλλη πλευρά, το τίμημα για την αντιμετώπιση κάθε πεδίου χωριστά κατά το συνδυασμό, είναι η αγνόηση περιπτώσεων με διφορούμενες προβλέψεις πεδίων από τα συστήματα του βασικού επιπέδου.

Η κατασκευή διανυσμάτων χαρακτηριστικών είναι μόνο ένας τρόπος χειρισμού των δεδομένων σε μετα-επίπεδο. Εναλλακτικοί τρόποι χρήσης των δεδομένων αυτών μπορούν επίσης να διερευνηθούν. Μια ενδιαφέρουσα λύση θα ήταν η κατάλληλη κωδικοποίηση της πληροφορίας αυτής ως ειδικές ετικέτες, οι οποίες θα μπορούσαν είτε να ενσωματωθούν μέσα στο κείμενο είτε να χρησιμοποιηθούν ως επιπλέον χαρακτηριστικά ορισμένων λεκτικών μονάδων μέσα στο κείμενο. Κάτι τέτοιο θα συμβάδιζε με δύο κύρια χαρακτηριστικά της συσσωρευμένης γενίκευσης, όπως ορίστηκαν από τον Wolpert [120] για προβλήματα κοινής ταξινόμησης. Το πρώτο χαρακτηριστικό είναι ότι τα δεδομένα τόσο σε βασικό όσο και σε μετα-επίπεδο είναι του ίδιου μεγέθους. Στην περίπτωση της εξαγωγής πληροφορίας δηλαδή, θα υπάρχει ο ίδιος αριθμός κειμένων, επισημειωμένων με την επιθυμητή πληροφορία, και στο βασικό και στο μετα-επίπεδο. Το δεύτερο χαρακτηριστικό αποτελεί άμεση συνέπεια του πρώτου και αφορά το γεγονός ότι ένας αλγόριθμος μάθησης σχεδιασμένος για εξαγωγή πληροφορίας θα μπορεί επίσης να εφαρμοστεί και σε μετα-επίπεδο.

Το πρόβλημα της εξαγωγής πληροφορίας θα μπορούσε επίσης να μοντελοποιηθεί ως ένα πρόβλημα *ακολουθιακής μάθησης (sequence learning)* σε μετα-επίπεδο. Με οδηγό το συσσωρευμένο σχεδιότυπο και δοθέντων των ακολουθιών των προβλέψεων των συστημάτων σε βασικό επίπεδο, στόχος θα είναι η εύρεση της ακολουθίας των σωστών πεδίων. Η εργασία [18] αποτελεί μια αρχική προσέγγιση προς την κατεύθυνση αυτή.

Ένα σημαντικό θέμα στην εξαγωγή πληροφορίας αφορά το υψηλό κόστος επισημείωσης αρχείων κειμένου για μια θεματική περιοχή. Πρόσφατη έρευνα στο χώρο της *ενεργής μάθησης (active learning)* έχει ως στόχο την επιλεκτική επιλογή σελίδων τις οποίες θα κληθεί να επισημειώσει ο ειδικός της περιοχής [83, 108]. Λιγότερες, αλλά σωστότερα επιλεγμένες, επισημειωμένες σελίδες μπορούν να οδηγήσουν σε ισοδύναμα αποτελέσματα εξαγωγής σε σχέση με περισσότερες σελίδες, αλλά τυχαία επιλεγμένες. Ο συνδυασμός συστημάτων εξαγωγής πληροφορίας δεν έχει μελετηθεί μέχρι τώρα για τη μείωση του κόστους επισημείωσης.

Οι τεχνικές ψηφοφορίας και συσσωρευμένης γενίκευσης που αξιολογήθηκαν σε αυτή τη διατριβή έχουν άμεση εφαρμογή στο χώρο της *μηχανικής οντολογιών (ontology engineering)*, ερχόμενοι περισσότερο κοντά στην εκπλήρωση των στόχων του σημασιολογικού ιστού για δημιουργία περιεχομένου κατανοητού από τις μηχανές των

υπολογιστών. Για παράδειγμα, εργασίες όπως η συντήρηση και εμπλουτισμός οντολογιών [113] με παραδείγματα, μπορούν να πραγματοποιηθούν με συνδυασμό πολλαπλών συστημάτων εξαγωγής πληροφορίας. Οι προτεινόμενες τεχνικές συνδυασμού μπορούν επίσης να συνδυαστούν με τις οντολογίες στην εργασία [34] για την εξαγωγή πληροφορίας από σελίδες του ιστού. Για παράδειγμα, οι χειρονακτικά κατασκευασμένοι κανόνες για την αναγνώριση παραδειγμάτων σχετικών πεδίων που θα εισαχθούν στη συνέχεια σε μια οντολογία, θα μπορούσαν να αντικατασταθούν από περισσότερο πολύπλοκους κανόνες, η εκμάθηση των οποίων θα προέκυπτε από το συνδυασμό συστημάτων. Κάτι τέτοιο θα ελάττωνε το κόστος κατασκευής και συντήρησης μιας οντολογίας.

Μεθοδολογίες συνδυασμού μπορούν να μελετηθούν επίσης και για το πρόβλημα της *αναγνώρισης σημασιολογικών ρόλων (semantic role identification)* στα συστατικά προτάσεων ελευθέρου κειμένου. Το FrameNet [8] και το ProBank [63] αποτελούν μέχρι στιγμής τις σημαντικότερες συλλογές κειμένων της Αγγλικής, επισημειωμένων με σημασιολογικούς ρόλους στα συστατικά των προτάσεων. Στο απλοϊκό παράδειγμα “On May 26th, Annie rode a donkey on the beach”, το τμήμα “Annie” αντιστοιχεί στο ρόλο “driver”, το τμήμα “a donkey” αντιστοιχεί στο ρόλο “vehicle”, ενώ το τμήμα “on the beach” αντιστοιχεί στο ρόλο “area”. Ολόκληρη η πρόταση αντιστοιχεί στο *σημασιολογικό πλαίσιο (semantic frame)* “transportation”.

Τα προβλήματα της εξαγωγής πληροφορίας και της αναγνώρισης σημασιολογικών ρόλων διαφέρουν αρκετά, απαιτώντας διαφορετικές προσεγγίσεις για την επίλυσή τους. Στην εξαγωγή πληροφορίας απαιτείται η αναγνώριση ενός πολύ μικρού αριθμού ρόλων, σχετικών με μια θεματική περιοχή. Αντίθετα, οι συλλογές FrameNet και ProBank δεν ανήκουν σε μια συγκεκριμένη περιοχή και απαιτούν την αναγνώριση σημαντικά μεγαλύτερου αριθμού ρόλων. Επίσης, η εξαγωγή πληροφορίας πραγματοποιείται σε επίπεδο ολόκληρης της σελίδας κειμένου, ενώ η αναγνώριση σημασιολογικών ρόλων σε μια σελίδα κειμένου πραγματοποιείται σε επίπεδο πρότασης. Μεθοδολογίες συνδυασμού αλγορίθμων που ήδη υπάρχουν για την αναγνώριση σημασιολογικών ρόλων (ενδεικτικά, [56, 109]) μπορούν να μελετηθούν.

Αισιοδοξώ ότι αυτή η διατριβή συνεισφέρει στην αναγνώριση της μεγάλης δυναμικής των μεθόδων συνδυασμού προς την κατεύθυνση του εντοπισμού σχετικής πληροφορίας στον τεράστιο όγκο δεδομένων κειμένου που είναι διαθέσιμα, προσδοκώντας μια μέθοδο εύκολα προσαρμόσιμη σε νέες θεματικές περιοχές.

ΠΑΡΑΡΤΗΜΑ Α: Πλήρη Πειραματικά Αποτελέσματα

Α.1 Σύγκριση ανά πεδίο σε βασικό επίπεδο με βάση την ακρίβεια

ΠΕΔΙΟ	BWI	HMM	$(LP)^2$
crsNumber	88.72	85.08	93.88
crsTitle	84.26	58.60	82.14
crsInst	55.50	48.16	53.33
projMember	60.43	59.03	64.80
projTitle	43.48	28.83	47.26
batterylife	100	79.17	100
batterytype	68.00	62.69	66.22
cdromspeed	83.33	69.09	88.10
dvdspeed	64.52	49.02	84.21
HDcapacity	76.22	80.31	63.46
manuf	85.00	55.24	61.43
model	47.67	45.83	35.29
modemSpeed	89.74	85.38	89.74
preinstOS	67.74	67.61	73.04
preinstSW	64.66	28.23	58.33
price	80.52	31.93	53.70
procName	70.59	66.50	58.19
procSpeed	78.57	77.55	51.63
ram	66.94	58.64	47.50
screenRes	96.67	88.89	83.33
screenSize	83.03	82.58	77.60
screenType	72.78	77.37	71.43
warranty	70.00	81.82	71.43
weight	100	100	86.21
application	83.77	75.28	80.24
area	78.74	47.91	77.31
city	97.30	95.42	97.86
company	91.52	49.64	89.23
country	88.52	74.19	99.34
desired_degree	26.32	09.30	57.14
desired_years_experience	59.09	48.72	93.18
Id	98.70	98.39	99.03
language	91.35	79.13	85.46
platform	86.61	71.69	77.89
post_date	98.02	91.40	98.67
recruiter	88.72	60.93	94.44
req_degree	92.41	84.78	94.05
req_years_experience	81.53	61.22	86.71
salary	91.21	69.35	95.87
state	97.04	96.86	98.09
title	75.97	65.95	69.44
etime	93.69	51.62	97.71
location	93.40	84.53	87.53
speaker	79.80	75.26	80.23
stime	100.00	99.09	99.33

Α.2 Σύγκριση ανά πεδίο σε βασικό επίπεδο με βάση την ανάκληση

ΠΕΔΙΟ	BWI	HMM	$(LP)^2$
crsNumber	71.49	87.19	95.04
crsTitle	40.27	55.75	61.06
crsInst	22.61	45.21	43.99
projMember	67.32	72.32	57.27
projTitle	9.90	31.68	34.16
batterylife	66.67	79.17	66.67
batterytype	47.89	59.15	69.01
cdromspeed	75.00	95.00	92.50
dvdspeed	52.63	65.79	42.11
HDcapacity	71.43	88.57	75.43
manuf	43.31	36.94	54.78
model	18.14	34.07	21.24
modemSpeed	89.74	94.87	89.74
preinstOS	48.09	73.28	64.12
preinstSW	54.74	51.09	61.31
price	32.63	40.00	45.79
procName	43.05	59.64	46.19
procSpeed	50.46	69.72	58.26
Ram	46.82	74.57	43.93
screenRes	67.44	93.02	69.77
screenSize	70.98	76.17	73.58
screenType	83.97	94.23	89.74
warranty	46.67	60.00	66.67
weight	86.21	62.07	86.21
application	58.60	55.43	76.63
area	37.16	60.75	55.24
city	94.45	93.07	98.46
company	65.37	60.17	75.32
country	97.34	98.34	100
desired_degree	23.81	19.05	19.05
desired_years_experience	88.64	86.36	93.18
Id	99.02	100	100
language	83.39	81.74	79.62
platform	72.08	75.93	75.10
post_date	99.66	96.31	99.66
recruiter	80.82	84.93	87.33
req_degree	82.95	88.64	89.77
req_years_experience	70.33	82.42	75.27
salary	56.08	87.16	78.38
state	97.04	96.70	98.09
title	35.96	73.21	47.52
etime	96.65	98.24	97.54
location	64.94	71.43	69.26
speaker	48.32	57.00	67.65
stime	96.46	99.09	97.61

Α.3 Σύγκριση ανά πεδίο σε βασικό επίπεδο με βάση τη μετρική F1

ΠΕΔΙΟ	BWI	HMM	$(LP)^2$
crsNumber	79.18	86.12	94.46
crsTitle	54.49	57.14	70.05
crsInst	32.13	46.64	48.21
projMember	63.69	65.00	60.80
projTitle	16.13	30.19	39.66
batterylife	80.00	79.17	80.00
batterytype	56.20	60.87	67.59
cdromspeed	78.95	80.00	90.24
dvdspeed	57.97	56.18	56.14
HDcapacity	73.75	84.24	68.93
manuf	57.38	44.27	57.91
model	26.28	39.09	26.52
modemSpeed	89.74	89.88	89.74
preinstOS	56.25	70.33	68.29
preinstSW	59.29	36.36	59.79
price	46.44	35.51	49.43
procName	53.48	62.88	51.50
procSpeed	61.45	73.43	54.74
Ram	55.10	65.65	45.65
screenRes	79.45	90.91	75.95
screenSize	76.54	79.25	75.53
screenType	77.98	84.97	79.55
warranty	56.00	69.23	68.97
weight	92.59	76.60	86.21
application	68.96	63.85	78.39
area	50.49	53.57	64.44
city	95.86	94.23	98.16
company	76.26	54.40	81.69
country	92.72	84.57	99.67
desired_degree	25.00	12.50	28.57
desired_years_experience	70.91	62.30	93.18
Id	98.86	99.19	99.51
language	87.19	80.42	82.44
platform	78.68	73.75	76.47
post_date	98.84	93.79	99.17
recruiter	84.59	70.96	90.75
req_degree	87.43	86.67	91.86
req_years_experience	75.52	70.26	80.59
salary	69.46	77.25	86.25
state	97.04	96.78	98.09
title	48.82	69.39	56.43
etime	95.15	67.68	97.62
location	76.61	77.43	77.33
speaker	60.20	64.87	73.41
stime	98.20	99.09	98.46

Α.4 Σύγκριση ανά πεδίο σε μετα-επίπεδο με βάση την ακρίβεια

ΠΕΔΙΟ	MVotM	MVotF	PVotM	PVotF	Stacking Nominal	Stacking Probs
crsNumber	94.38	94.30	94.38	94.76	92.74	93.09
crsTitle	73.01	91.67	73.99	77.66	77.84	84.27
crsInst	42.30	67.62	42.24	55.79	70.05	67.37
projMember	50.49	69.07	50.52	65.49	69.52	78.01
projTitle	32.02	60.00	32.17	36.81	52.70	86.84
batterylife	80.00	100	80.00	90.00	100	100
batterytype	64.21	65.45	64.21	76.32	69.64	72.73
cdromspeed	70.37	86.05	70.37	86.36	88.10	88.10
dvdspeed	45.31	90.00	45.31	74.19	78.95	68.97
HDcapacity	65.70	82.66	68.09	82.70	81.91	90.80
manuf	51.71	88.73	52.86	83.50	85.19	95.65
model	37.26	55.81	37.31	44.83	60.00	65.28
modemSpeed	85.38	89.74	85.38	89.17	89.74	89.74
preinstOS	56.63	80.20	56.99	76.15	76.00	77.60
preinstSW	28.73	90.24	28.85	66.92	85.90	90.79
price	30.50	85.54	30.95	49.45	64.41	85.71
procName	55.28	73.94	55.48	71.07	74.68	82.98
procSpeed	54.60	86.90	55.27	80.83	84.21	90.12
ram	46.90	72.73	46.78	63.82	69.70	74.64
screenRes	83.33	90.91	83.33	87.50	92.50	88.37
screenSize	71.36	85.21	71.04	83.70	88.68	89.29
screenType	70.89	74.87	70.89	74.63	77.96	75.82
warranty	67.74	72.00	67.74	95.00	72.00	100
weight	93.10	100	93.10	96.15	100	100
application	71.68	89.36	71.91	78.64	80.71	84.12
area	49.89	88.17	50.25	66.26	78.80	83.75
city	94.01	97.62	94.02	97.05	97.41	97.85
company	60.32	92.17	60.69	73.41	89.23	90.70
country	74.26	88.62	74.26	77.52	99.34	99.34
desired_degree	18.42	80.00	18.92	26.92	100	11.11
desired_years_experience	51.25	88.64	51.35	55.07	97.44	91.18
Id	98.39	98.71	98.39	99.02	99.03	99.03
language	78.24	93.29	78.31	84.54	91.15	89.90
platform	67.14	90.97	66.53	74.19	90.83	83.29
post_date	91.95	98.02	91.95	95.16	98.67	99.33
recruiter	63.45	90.88	63.97	73.10	94.44	93.73
req_degree	81.31	95.18	81.31	85.87	93.10	90.00
req_years_experience	63.84	88.05	61.51	70.46	88.34	89.68
salary	67.29	100	67.29	72.19	96.69	80.54
state	95.63	97.73	95.79	96.90	97.77	98.06
title	61.20	87.90	61.78	70.54	71.84	81.92
etime	81.50	98.41	81.86	95.85	97.72	98.75
location	85.19	96.42	85.19	87.80	90.44	91.57
speaker	77.03	93.29	77.03	79.03	81.73	87.85
stime	99.51	100	99.51	99.51	99.67	100

Α.5 Σύγκριση ανά πεδίο σε μετα-επίπεδο με βάση την ανάκληση

ΠΕΔΙΟ	MVotM	MVotF	PVotM	PVotF	Stacking Nominal	Stacking Probs
crsNumber	97.11	75.21	97.11	97.11	95.04	94.63
crsTitle	73.01	48.67	73.01	67.70	60.62	66.37
crsInst	63.75	33.60	63.75	59.88	28.11	51.73
projMember	84.22	72.09	84.22	72.77	68.50	69.90
projTitle	36.14	11.88	36.63	29.70	19.31	16.34
batterylife	83.33	66.67	83.33	75.00	66.67	66.67
batterytype	85.92	50.70	85.92	81.69	54.93	56.34
cdromspeed	95.00	92.50	95.00	95.00	92.50	92.50
dvdspeed	76.32	47.37	76.32	60.53	39.47	52.63
HDcapacity	90.86	81.71	91.43	87.43	88.00	84.57
manuf	67.52	40.13	70.70	54.78	43.95	56.05
model	43.36	21.24	44.25	34.51	13.27	20.80
modemSpeed	94.87	89.74	94.87	91.45	89.74	89.74
preinstOS	84.73	61.83	83.97	75.57	58.02	74.05
preinstSW	75.91	54.01	76.64	63.50	48.91	50.36
price	54.74	37.37	54.74	47.37	40.00	44.21
procName	70.40	47.09	70.40	62.78	52.91	52.47
procSpeed	78.90	57.80	79.36	71.56	66.06	71.10
ram	78.61	55.49	79.77	73.41	66.47	59.54
screenRes	93.02	69.77	93.02	81.40	86.05	88.37
screenSize	81.35	74.61	81.35	79.79	73.06	77.72
screenType	96.79	91.67	96.79	96.15	92.95	74.36
warranty	70.00	60.00	70.00	63.33	60.00	53.33
weight	93.10	86.21	93.10	86.21	79.31	86.21
application	81.14	63.11	82.47	81.14	75.46	76.96
area	75.96	46.86	76.07	65.82	56.56	59.65
city	99.23	94.61	99.38	96.46	98.61	98.00
company	82.25	66.23	83.55	80.09	75.32	67.53
country	99.67	98.34	99.67	99.67	99.67	100
desired_degree	33.33	19.05	33.33	33.33	04.76	04.76
desired_years_experience	93.18	88.64	86.36	86.36	86.36	70.45
Id	100	100	100	98.69	100	100
language	93.99	83.51	93.99	90.81	84.92	87.04
platform	91.88	76.20	91.33	88.58	76.34	79.50
post_date	99.66	99.66	99.66	98.99	99.66	99.66
recruiter	94.52	85.27	94.86	92.12	87.33	86.99
req_degree	98.86	89.77	98.86	89.77	92.05	92.05
req_years_experience	95.05	76.92	93.96	91.76	79.12	76.37
salary	97.30	75.00	97.30	82.43	79.05	81.08
state	98.96	97.22	98.96	97.91	98.96	96.52
title	86.24	45.32	86.61	72.48	67.89	64.04
etime	97.71	98.06	97.71	97.54	98.06	97.71
location	78.70	69.96	78.70	76.62	74.55	77.14
speaker	74.75	53.45	74.75	72.09	71.01	71.99
stime	99.51	99.09	99.51	99.51	99.67	100

Α.6 Σύγκριση ανά πεδίο σε μετα-επίπεδο με βάση τη μετρική F1

ΠΕΔΙΟ	MVotM	MVotF	PVotM	PVotF	Stacking Nominal	Stacking Probs
crsNumber	95.72	83.68	95.72	95.92	93.88	93.85
crsTitle	73.01	63.58	73.50	72.34	68.16	74.26
crsInst	50.85	44.90	50.81	57.76	40.12	58.53
projMember	63.13	70.55	63.16	68.94	69.00	73.73
projTitle	33.95	19.83	34.26	32.88	28.26	27.50
batterylife	81.63	80.00	81.63	81.82	80.00	80.00
batterytype	73.49	57.14	73.49	78.91	61.42	63.49
cdromspeed	80.85	89.16	80.85	90.48	90.24	90.24
dvdspeed	56.86	62.07	56.86	66.67	52.63	59.70
HDcapacity	76.26	82.18	78.05	85.00	84.85	87.57
manuf	58.56	55.26	60.49	66.15	57.98	70.68
model	40.08	30.77	40.49	39.00	21.74	31.54
modemSpeed	89.88	89.74	89.88	90.30	89.74	89.74
preinstOS	67.89	69.83	67.90	75.86	65.80	75.78
preinstSW	41.68	67.58	41.92	65.17	62.33	64.79
price	39.17	52.01	39.54	48.39	49.35	58.33
procName	61.93	57.53	62.06	66.67	61.94	64.29
procSpeed	64.54	69.42	65.16	75.91	74.04	79.49
ram	58.75	62.95	58.97	68.28	68.05	66.24
screenRes	87.91	78.95	87.91	84.34	89.16	88.37
screenSize	76.03	79.56	75.85	81.70	80.11	83.10
screenType	81.84	82.42	81.84	84.03	84.80	75.08
warranty	68.85	65.45	68.85	76.00	65.45	69.57
weight	93.10	92.59	93.10	90.91	88.46	92.59
application	76.12	73.97	76.83	79.87	78.00	80.38
area	60.23	61.20	60.53	66.04	65.85	69.67
city	96.55	96.09	96.63	96.75	98.01	97.92
company	69.60	77.08	70.31	76.60	81.69	77.42
country	85.11	93.23	85.11	87.21	99.50	99.67
desired_degree	23.73	30.77	24.14	29.79	09.09	06.67
desired_years_experience	66.13	88.64	64.41	67.26	91.57	79.49
Id	99.19	99.35	99.19	98.85	99.51	99.51
language	85.39	88.13	85.44	87.56	87.93	88.45
platform	77.58	82.93	76.99	80.75	82.96	81.35
post_date	95.65	98.84	95.65	97.04	99.17	99.50
recruiter	75.93	87.99	76.41	81.52	90.75	90.23
req_degree	89.23	92.40	89.23	87.78	92.57	91.01
req_years_experience	76.38	82.11	74.35	79.71	83.48	82.49
salary	79.56	85.71	79.56	76.97	86.99	80.81
state	97.26	97.47	97.35	97.40	98.36	97.28
title	71.59	59.81	72.12	71.49	69.81	71.88
etime	88.87	98.24	89.09	96.68	97.89	98.23
location	81.82	81.08	81.82	81.83	81.73	83.74
speaker	75.88	67.96	75.88	75.40	75.99	79.13
stime	99.51	99.55	99.51	99.51	99.67	100

Συγκρίνοντας τους Πίνακες Α.3 και Α.6, παρατηρείται μείωση στο F1 για μερικά πεδία από τη συσσωρευμένη γενίκευση με πιθανότητες σε σχέση με το F1 του καλύτερου συστήματος σε βασικό επίπεδο. Μόνο σε ένα πεδίο η ελάττωση αυτή μετρήθηκε ως στατιστικά σημαντική.

Α.7 Σύγκριση μεθόδων συνδυασμού σε ζευγάρια συστημάτων εξαγωγής πληροφορίας

		BWI+HMMs			BWI+(LP)^2			HMMs+(LP)^2			BWI+HMMs+(LP)^2		
		P	*R.*	*F1*	*P*	*R.*	*F1*	*P*	*R.*	*F1*	*P*	*R.*	*F1*
Μαθήματα	*MVotM*	57.93	62.46	60.25	69.84	62.77	66.12	59.63	73.62	65.89	58.68	74.35	65.59
	MVotF	60.62	57.14	58.83	70.03	47.76	56.79	77.27	37.23	50.25	82.05	47.65	60.29
	PVotM	57.93	62.46	60.11	69.84	62.77	66.12	59.76	73.73	66.01	58.78	74.35	65.65
	PVotF	65.84	55.27	60.09	81.37	60.58	69.46	72.64	70.59	71.60	70.16	71.12	70.64
	Stacking	76.00	43.59	55.40	78.02	52.55	62.80	86.63	50.68	63.95	81.32	52.66	63.92
	Stacking Probs	74.43	47.65	58.11	76.97	61.00	68.06	77.65	64.86	70.68	79.03	66.01	71.93
Προγράμματα	*MVotM*	51.46	73.83	60.65	55.67	73.58	63.38	53.29	78.82	63.59	49.17	79.32	60.71
	MVotF	58.25	65.36	61.60	58.66	68.18	63.06	55.25	64.20	59.39	68.88	65.96	67.39
	PVotM	51.46	73.83	60.65	55.63	73.47	63.32	53.29	78.77	63.57	49.20	79.37	60.75
	PVotF	61.82	59.61	60.69	85.54	31.62	46.17	63.01	64.95	63.97	63.31	68.38	65.75
	Stacking	70.42	55.47	62.06	60.72	58.14	59.40	73.12	50.63	59.83	68.84	63.49	66.05
	Stacking Probs	70.41	54.01	61.13	73.22	61.22	66.69	89.13	53.76	67.07	78.22	64.45	70.67
Φορητοί	*MVotM*	60.87	68.11	64.29	65.25	64.61	64.93	56.82	70.94	63.10	52.89	76.00	62.37
	MVotF	81.18	54.20	65.00	69.87	59.76	64.42	58.60	67.40	62.69	80.41	59.05	68.09
	PVotM	60.84	68.07	64.25	65.33	64.70	65.01	57.43	71.70	63.78	53.21	76.47	62.76
	PVotF	72.20	62.42	66.95	81.37	57.28	67.23	78.01	65.54	71.23	72.86	69.30	71.03
	Stacking	79.35	58.67	67.46	73.25	53.48	61.82	78.18	57.28	66.11	79.52	60.10	68.46
	Stacking Probs	80.65	57.66	67.24	84.53	54.41	66.20	82.99	61.32	70.53	84.49	62.04	71.55
Αγγελίες	*MVotM*	73.70	84.96	78.93	84.44	82.99	83.71	72.51	89.63	80.17	71.29	90.88	79.90
	MVotF	80.66	76.46	78.50	84.61	83.11	83.86	74.64	86.75	80.24	93.06	76.31	83.85
	PVotM	73.83	85.11	79.07	84.34	82.89	83.61	72.72	89.90	80.40	71.37	90.98	79.99
	PVotF	82.72	77.71	80.14	87.15	77.11	81.82	81.04	83.90	82.44	80.08	86.45	83.15
	Stacking	88.27	75.44	81.35	87.80	50.57	84.03	86.88	81.07	83.87	89.89	81.82	85.67
	Stacking Probs	89.33	75.92	82.08	89.27	79.58	84.14	88.78	81.31	84.88	90.27	82.00	85.94
Σεμινάρια	*MVotM*	86.52	83.17	84.81	90.20	84.03	87.01	87.24	86.16	86.70	86.93	86.82	86.87
	MVotF	91.75	74.54	82.26	95.89	71.96	82.22	95.53	71.43	81.74	97.55	78.72	87.13
	PVotM	87.94	84.29	86.07	90.22	84.03	87.02	87.31	86.16	86.73	86.99	86.82	86.90
	PVotF	93.07	80.92	86.57	91.13	83.45	87.12	91.48	84.49	87.85	90.69	85.50	88.02
	Stacking	89.79	82.57	86.03	90.83	82.95	86.71	90.46	84.21	87.22	92.57	84.74	88.48
	Stacking Probs	92.38	80.11	85.81	94.42	82.21	87.89	94.50	84.77	89.37	94.69	85.80	90.03

ΠΑΡΑΡΤΗΜΑ Β: Πλήρη Συγκριτικά Αποτελέσματα

Β.1 Σύγκριση μεθόδων συνδυασμού με καθολική συμφωνία στις προβλέψεις των συστημάτων του βασικού επιπέδου

	Συνολικό	Ψηφοφορία		Συσσωρευμένη Γενίκευση	
		MVotM, MVotF, PVotM	*PVotF*	*Ονομαστικές τιμές*	*Πιθανοτικές τιμές*
Μαθήματα	188	177	177	177	179
Προγράμματα	913	727	731	727	768
Φορητοί	1112	977	980	977	960
Αγγελίες	4119	4034	4034	4034	4034
Σεμινάρια	2041	2032	2032	2032	2033

Β.2 Σύγκριση μεθόδων συνδυασμού με μερική συμφωνία στις προβλέψεις των συστημάτων βασικού επιπέδου

	Σύνολο	Ψηφοφορία				Συσσωρευμένη γενίκευση	
		MVotM	*MVotF*	*PVotM*	*PVotF*	*Ονομαστικές τιμές*	*Πιθανοτικές τιμές*
Μαθήματα	1299	641	726	640	822	814	885
Προγράμματα	2419	931	1696	931	1544	1668	1857
Φορητοί	2168	749	1583	749	1548	1569	1724
Αγγελίες	4030	1782	1874	1782	2823	2943	2965
Σεμινάρια	2232	1339	1366	1337	1371	1417	1529

Β.3 Σύγκριση μεθόδων συνδυασμού με διαφωνία στις προβλέψεις των συστημάτων του βασικού επιπέδου

	Σύνολο	Ψηφοφορία				Συσσωρευμένη γενίκευση	
		MVotM	*MVotF*	*PVotM*	*PVotF*	*Ονομαστικές τιμές*	*Πιθανοτικές τιμές*
Μαθήματα	7	4	4	5	5	3	6
Προγράμματα	5	0	1	1	1	2	3
Φορητοί	127	76	78	87	89	99	110
Αγγελίες	471	329	325	336	336	340	346
Σεμινάρια	423	414	417	416	416	417	419

ΑΝΑΦΟΡΕΣ

1. Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P., Vilain, M., MITRE: Description of the Alembic system used for MUC-6. *In Proceedings of the 6th Message Understanding Conference (MUC)*, 141-155, Morgan Kaufmann, 1995.
2. Adelberg, B., NoDoSE: A tool for semi-automatically extracting structured and semi-structured data from text documents. *In Proceedings of the ACM SIGMOD International Conference on Management of Data*, Seattle, USA, 283-294, 1998.
3. Agrawal, R., Srikant, R., Fast algorithms for mining association rules in large databases. *In Proceedings of the International Conference in very large databases (VLDB)*, Morgan Kaufmann, 478-499, 1994.
4. Aha, D., Kibler, D., Instance-based learning algorithms, *Machine Learning*, 6, 37-66, 1991.
5. Ali, K. M., Pazzani, M. J., Error reduction through learning multiple descriptions. *Machine Learning*, 24, 173-202, 1996.
6. Al-Ani, A., Deriche, M., A new technique for combining multiple classifiers using the Dempster-Shafer theory of evidence, *Journal of Machine Learning Research (JMLR)*, 17, 333-361, 2002.
7. Arocena, G.O., Mendelzon, A.O., WebOQL: Restructuring documents, databases, and webs. *In Proceedings of the 14th International Conference on Data Engineering*, 24-33, Orlando, USA, 1998.
8. Baker, C. F., Fillmore, C. J., Lowe, J. B., The Berkeley FrameNet Project, *In Procedings of the COLING-ACL' 98 conference,* 1998.
9. Bauer, E., Kohavi, R., An empirical comparison of voting classification algorithms, *Machine Learning*, 36(1-2), 105-139, 1999.
10. Bikel, D., Schwartz, R., Weischedel, R., An algorithm that learns what's in a name, *Machine Learning*, *Special Issue on Natural Language Learning*, 34(1-3), 1999.
11. Brazdil, P., Soares, C., A Comparison of Ranking Methods for Classification Algorithm Selection, *In Proceedings of the 11th European Conference on Machine Learning (ECML)*, LNAI 1810, Springer Verlag, 2000.
12. Brazdil, P., Gama, J., Henery, B., Characterizing the applicability of classification algorithms using meta-level learning. *In Proceedings of the 7th European Conference on Machine Learning* (ECML), Catania, Italy, 83-102, 1994.
13. Breiman L., Bagging Predictors, *Machine Learning,* 24(2), 123-140, 1996.
14. Breiman L., Stacked Regressions, *Machine Learning*, 24, 41-48, 1996a.
15. Brill, E., A corpus-based approach to language learning, *PhD dissertation*, University of Pennsylvania, 1993.
16. Califf M.E., Mooney R.J., Bottom-up Relational Learning of Pattern Matching Rules for Information Extraction, *Journal of Machine Learning Research*, 4, 177-210, 2003.
17. Cardie, C., Empirical methods in information extraction, *AI Magazine*, Volume 18(4), 65-79, 1997.
18. Carvalho, V. R., Cohen, W., Stacked Sequential Learning, *In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI),* 2005.
19. Chan, P., Stolfo, S., Experiments on multistrategy-learning by meta-learning, *In Proceedings of the 2nd International Conference on Information and Knowledge Management (CIKM)*, 314-323, 1993.
20. Chang, C.H., Lui, S.C., IEPAD : Information Extraction based on Pattern Discovery, *In Proceedings of the 10th WWW conference*, 509-516, New York, USA, 2001.
21. Chawathe, S., Molina, H-C., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J., Widom, J., The TSIMMIS Project: Integration of Heterogeneous Information Sources, *In Proceedings of the 10th Meeting of Information Processing Society of Japan (IPSJ)*, 7-18, 1994.
22. Cheeseman, P., Stutz, J., Bayesian classification (AutoClass): Theory and results. *Advances in Knowledge Discovery and Data Mining*, AAAI Press, 153-180, 1995.

23. Cherkauer, K.J., Human expert-level performance on a scientific image analysis task by a system using combined artificial networks, *Working Notes of the AAAI workshop on integrating multiple learned models,* 15-21, 1996.

24. Ciravegna, F., Adaptive Information Extraction from Text by Rule Induction and Generalization. *In Proceedings of the 17th IJCAI Conference.* Seattle, 2001.

25. Ciravegna, F., Dingli, A., Petrelli, D., Wilks, Y., User-system Cooperation in Document Annotation based on Information Extraction, *In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, Spain, October, 2002.

26. Ciravegna, F., Lavelli, A., LearningPinochio: Adaptive Information Extraction for Real World Applications, *Natural Language Engineering* 1(1), 1-21, Cambridge University Press, 2003.

27. Ciravegna, F., Wilks, Y., Designing Adaptive Information Extraction for the Semantic Web in Amilcare, *In Handschuh and Staab (eds), Annotation for the Semantic Web, in the Series Frontiers in Artificial Intelligence and Applications*, IOS Press. 2003.

28. Cohen, W., Fan, W., Learning Page-Independent Heuristics for Extracting Data from Web Pages. *In Proceedings of the 8th International World Wide Web (WWW) Conference*, Toronto, Canada, 1999.

29. Cohen, W., Hurst, M., Jensen, L.S., A Flexible Learning System for Wrapping Tables and Lists in HTML Documents. *In Proceedings of the 11th International World Wide Web conference (WWW)*, Hawaii, USA, 2002.

30. Craven, M., DiPasquo, D., Freitag, D., McCallum, A.K., Mitchell, T., Nigam, K., Slattery, S., Learning to extract symbolic knowledge from the World Wide Web. *In Proceedings of the 15th National Conference on Artificial Intelligence (AAAI),* 1998.

31. Crescenzi, V., Mecca, G., Grammars have exceptions, *Information Systems*, 23(8), 539-565, 1998.

32. Crescenzi, V., Mecca, G., Merialdo, P., RoadRunner: Towards automatic data extraction from large Web sites. *In Proceedings of the 16th International Conference on Very Large Data Bases (VLDB)*, 109-118, Rome, Italy, 2001.

33. Cross-Lingual multi-agent retail comparison (CROSSMARC), *R&D project*, Institute of Informatics and Telecommunications, National Center of Scientific Research (NCSR) “Demokritos”, Athens, Greece, http://www.iit.demokritos.gr/skel/crossmarc/., 2003.

34. Davulcu, H., Mukherjee, S., Ramakrishman, I.V., Extraction Techniques for Mining Services from Web Sources, *IEEE International Conference on Data Mining*, Maebashi City, Japan, 2002.

35. Defense Advanced Research Projects Agency (DARPA), *Proceedings of the 6th Message Understanding Conferences (MUC-6)*, Morgan Kaufmann, 1995.

36. Defense Advanced Research Projects Agency (DARPA), *Proceedings of the 7th Message Understanding Conferences (MUC-7)*, Morgan Kaufmann, 1996.

37. Dietterich, T.G., Machine Learning research: Four current directions. *AI Magazine*, 18(4), 97-136, 1997.

38. Dietterich T.G., Approximate Statistical Tests for Comparing Supervised Machine Learning Algorithms, *Neural Computing,*10(7), 1895-1924, 1998.

39. Domingos, P., Unifying instance-based and rule-based induction. *Machine Learning*, 24(2), 141-168, 1996.

40. Doorenbos, R.B., Etzioni O., Weld, D.S., A scalable comparison shopping agent for the world wide web. *In Proceedings of the 1st International Conference on Autonomous Agents*, 1997.

41. Džeroski, S., Ženko, B., Is Combining Classifiers Better than Selecting the Best One? *Machine Learning*, 54(3): 255-273, 2004.

42. Embley, D.W., Campbell, D.M., Jiang, Y.S., Liddle, S.W., Lonsdale, D.W., Ng Y.K., Smith, R.D., Conceptual model-based data extraction from multiple-record web documents, *Data and Knowledge Engineering*, 31(3), 227-251, 1999.

43. Etzioni, O., The world wide web: Quagmire or gold mine. *Communications of the ACM*, 39(11), 65-68, 1996.

44. Fisher, D., Knowledge acquisition via incremental conceptual clustering. *Machine Learning,* 2(2), 139-172, 1987.

45. Florian, R., Named Entity Recognition as a House of Cards: Classifier Stacking, *In Proceedings of the 6th conference on Computational Natural Language Learning (CoNLL)*, 175-178, 2002.

46. Florian, R., Cucerzan, S., Schafer, C., Yarowsky, D., Combining Classifiers for Word Sense Disambiguation, *Natural Language Engineering*, 1(1), 1-14, 2002.

47. Frank, E., Wang, Y., Inglis, S., Holmes, G., Witten, I.E., Using model trees fro classification. *Machine Learning*, 32(1), 63-76, 1998.

48. Freitag, D., Machine Learning for Information Extraction in Informal Domains, *Machine Learning*, 39, 169-202, 2000.

49. Freitag, D., Kushmerick, N., Boosted Wrapper Induction, *In Proceedings of the 17th National conference on Artificial Intelligence.* (AAAI-1999), 59-66, 1999.

50. Freitag., D., McCallum, A.K., Information extraction with HMM structures learned by stochastic optimization, *In Proceedings of the 18th National conference on Artificial Intelligence* (AAAI-2000), 584-589, 2000.

51. Freund, Y., Schapire, R., Experiments with a new boosting algorithm. *In Proceedings of the 13th International Conference on Machine Learning (ICML)*, 148-156, 1996.

52. Friedman, J., Hastie, T., Tibshirani, R., Additive Logistic Regression: a Statistical View Of Boosting. *Technical Report*, Stanford University, 1999.

53. Gama, J., Brazdil, P., Characterization of classification algorithms, *In Proceedings of the 7th Portuguese Conference on Artificial Intelligence*, 189-200, 1995.

54. Gama, J., Brazdil, P., Cascade generalization, *Machine Learning*, 41(3), 315-344, 2000.

55. Ghani, R., Jones, R., Mladenić, D., Nigam, K., Slattery, S., Data mining on symbolic knowledge extracted from the Web, *In Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (KDD-2000), Workshop on Text Mining*, 29–36, Boston, MA, 2000.

56. Gildea, D., Jurafsky D., Automatic labeling of semantic roles, *Computational Linguistics*, 28, 245-288, 2002.

57. Halteren, H., Zavrel J., Daelemans, W., Improving Accuracy in Word Class Tagging through Combination of Machine Learning Systems, *Computational Linguistics*, 27(2), 199-230, 2001.

58. Hsu, C.N., Dung, M.T., Generating finite-state transducers for semi structured data extraction from the Web, *Information Systems*, *Special Issue on semi structured data*, 23(8), 1998.

59. Japkowicz, N., Stephen, S., The Class Imbalance Problem: A Systematic Study. *Intelligent Data Analysis*, 6(5), 429-450, November, 2002.

60. John, G.H., Langley, P., Estimating Continuous Distributions in Bayesian Classifiers, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. pp. 338-345. Morgan Kaufmann, San Mateo, 1995.

61. Kalousis, A., Gama, J., Hilario, M., On Data and Algorithms: Understanding Inductive Performance, *Machine Learning*, 54, 275-312, 2004.

62. Kauchak, D., Smarr, J., Elkan, C., Sources of Success for Boosted Wrapper Induction, *Journal of Machine Learning Research*, 5, 499-527, 2004.

63. Kingsbury, P., Palmer, M., Marcus, M., Adding semantic annotation to the penn Treebank, *In Proceedings of the Human Computer Interaction technology conference*, San Diego, USA, 2002.

64. Knoblock, C, Minton, S., Ambite, J.L., Ashish, N., Muslea, I., Philpot, A. G., Tejada, S. The ariadne approach to web-based information integration, *IEEE Intelligent Systems*, 13(5), 1998.

65. Kononenko, I., Kovačič, M., Learning as optimization: stochastic generation of multiple knowledge, *In Proceedings of the 9th International Conference on Machine Learning (ICML)*, pp. 257-262, 1992.

66. Kosala, R., Blockeel, H., Web mining research: a survey, *ACM SIGKDD Explorations Newsletter, 2(1)*, 1-15, 2000.

67. Krogh, A., Vedelsby, J., Neural Networks Ensembles, Cross Validation and Active Learning, *Advances in Neural Information Processing Systems,* 7, 231-238, MIT Press, 1995.

68. Kuncheva, L.I., Whitaker, C.J., Measures of Diversity in Classifier Ensembles and their Relationship with the Ensemble Accuracy, *Machine Learning*, 51, 181-207, 2003.

69. Kushmerick, N., Wrapper Induction for Information Extraction, *PhD Thesis*, University of Washington, 1997.

70. Kushmerick, N., Regression testing for wrapper maintenance. In Proceedings of the 16th National Conference on Artificial Intelligence (AAAI), 74-79, 1999.

71. Kushmerick, N., Thomas, N., Adaptive information extraction: Core technologies for information agents. *In Intelligent information agents R&D in Europe, LNAI 2586*, Springer, 2002.

72. Kwok, S.W., Carter, C., Multiple decision trees, *Uncertainty in Artificial Intelligence,* 4, 327-335, Elsevier Science, 1990.

73. Laender, A., Ribeiro-Neto B., da Silva A., Teixeira J., A Brief Survey of Web Data Extraction Tools, *SIGMOD Record*, 31(2), 2002.

74. Lafferty, J., McCallum, A., Pereira, F., Conditional random fields: Probabilistic models for segmenting and labelling sequence data, *In Proceedings of the 18th International Conference on Machine Learning (ICML)*, 2001.

75. Lavelli, A., Califf, M.E., Ciravegna, F., Freitag, D., Giuliano, C., Kushmerick, N., Romano, L., A critical survey of the methodology for IE evaluation, *In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC),* Lisbon, Portugal, 2004.

76. Liu, L., Pu, C., and Ilan, W. XWRAP: An XML-enabled wrapper construction system for web information sources. *In Proceedings of the International Conference on Data Engineering*, 2000.

77. Mecca, G., Atzeni, P., Masci, A., Merialdo, P., Sindoni, G., From Databases to Web-Bases: The ARANEUS Experience - *Technical Report n. 34-1998* - Dipartimento di Informatica e Automazione, Universita' di Roma Tre, May, 1998.

78. Merz, C., J., Using correspondence analysis to combine classifiers, *Machine Learning*, 36 (1), 33-58, 1999.

79. Michalski, R., Tecuci , G., (Eds) Machine learning: A multistrategy approach. *SanMateo, CA, Morgan Kaufmann*, 1994.

80. Mitchell, T.M., Machine Learning, *The McGraw-Hill Companies, Inc.*, 1997.

81. Muslea, I., Extraction patterns for information extraction tasks: A survey. *In Proceedings of AAAI 1999. Workshop on Machine Learning for Information Extraction*, 1999.

82. Muslea, I., Minton, S., Knoblock, C., Hierarchical Wrapper Induction for Semistructured Information Sources, *Journal Of Autonomous Agents and Multi-Agent Systems*, 4, 93-114, 2001.

83. Muslea, I., Active learning with multiple views, *PhD Dissertation*, *University of Southern California*, 2002.

84. Nahm, U.Y., Mooney, R.J., A mutually beneficial integration of data mining and information extraction, *In Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00)*, 627-632, 2000.

85. Oliver, J.J., Hand, D.J., On pruning and averaging decision trees, *In Proceedings of the 12th International Conference on Machine Learning (ICML)*, 430-437, 1995.

86. Perrone, M.P., Cooper, L.N., When networks disagree: Ensembles methods for hybrid neural networks, *In Mammone, R.J. (Ed.), Neural networks for speech and image processing*, Chapman and Hall, 1993.

87. Pfahringer, B., Bensusan, H., Giraud-Carrier, C., Meta-learning by landmarking various learning algorithms. *In Proceedings of the 17th International Conference on Machine Learning (ICML)*. Stanford, CA., 2000.

88. Platt, J., Fast Training of Support Vector Machines using Sequential Minimal Optimization. *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, eds., MIT Press, 1998.

89. Quinlan, R., J., Induction of decision trees, Machine Learning, 1(1): 81-106, 1986.

90. Quinlan, R., J., C4.5: Programs for Machine Learning, *Morgan Kaufmann*, 1993.

91. Rabiner, L., A tutorial on hidden Markov models and selected applications in speech recognition. *In Proceedings of the IEEE 37-2*, 1989.

92. Sahuguet, A., Azavant, F., Building intelligent Web applications using lightweight wrappers, *Data and Knowledge Engineering*, 36(3), 283-316, 2001.

93. Schaffer, C., Cross-validation, stacking and bi-level stacking: Meta-methods for classification learning. *In P. Cheeseman and R. W. Oldford (Eds.), Selecting models from data: Artificial Intelligence and Statistics IV*, 51-59, Springer-Verlag, 1994.

94. Schaffer, C., A conservation law of generalization performance, *In Proceedings of the 11th International Conference on Machine Learning (ICML)*, 259-265, 1994b.

95. Sebastiani, F., Machine Learning for Automated Text Categorization, *ACM Computing Surveys (CSUR)*, 34 (1), 1-47, 2002.

96. Seewald, A., Exploring the parameter state space of stacking. *In Proceedings of the IEEE International Conference on Data Mining (ICDM)*, Japan, 2002.

97. Seewald, A., How to make Staking Better and Faster While Also Taking Care of an Unknown Weakness, *In Proceedings of the 19th International Conference in Machine Learning (ICML)*, San Francisco, 2002a.

98. Seewald, A., Towards understanding stacking, *PhD Thesis*, Dept. of Informatics, Technical University of Wien, Austria, 2003.

99. Seewald, A., Fürnkranz, J., An evaluation of grading classifiers, *Advances in Intelligent Data Analysis (IDA)*, 115-124, 2001.

100. Seymore, K., McCallum A.K., Rosenfeld, R., Learning hidden Markov model structure for Information Extraction. *Journal of Intelligent Information Systems,* 8(1): 5-28, 1999.

101. Shafer, G., A mathematical theory of evidence, *Princeton University Press*, 1976.

102. Sigletos, G., Paliouras, G., Karkaletsis, E., Role identification from free-text using Hidden Markov models, *Methods and Applications of Artificial Intelligence*, LNAI 2308, Springer-Verlag, 2002.

103. Sigletos, G., Paliouras, G., Spyropoulos, C.D., Stamatopoulos, T., Stacked Generalization for Information Extraction, *In Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*, IOS Press, Valencia, Spain, 2004.

104. Sigletos, G., Paliouras, G., Spyropoulos, C.D., Hatzopoulos, M., Mining web sites using wrapper induction, named entities and post-processing, Web Mining: from web to semantic web, *In Proceedings of the 1st European Web Mining Forum*, LNAI 3209, Springer, 97-112, 2004.

105. Smyth, P., Wolpert, D., Stacked Density Estimation. *Advances in Neural Information Processing Systems*, 1997.

106. Sonderland, S., Learning Information Extraction Rules for Semi-structured and Free Text, *Machine Learning*, 34-(1/3), 233-272, 1999.

107. Ting, K., Witten M., Issues in stacked generalization, *Journal of Artificial Intelligence Research (JAIR)*, 10, 271-289, 1999.

108. Thompson, C.A., Califf, M.E., Mooney, R.J., Active Learning for Natural Language Parsing and Information Extraction, *In Proceedings of the 16th International Machine Learning Conference (ICML)*, Bled, Slovenia, 1999.

109. Thompson, C.A., Levy, R., Manning, C.D., A generative model fro semantic role labelling, *In Proceedings of the 14th European Conference on Machine Learning (ECML)*, Cavtat-Dubrovnik, Croatia, 2003.

110. Todorovski, L., Džeroski, S., Combining classifiers with meta decision trees. *Machine Learning*, 50 (3), 223-249, 2002.

111. Todorovski, L., Blockeel, H., Džeroski, S., Ranking with predictive clustering trees. *In Proceedings of the 13th European Conference on Machine Learning (ECML)*, 444-455, 2002.

112. Tsoumakas, G., Katakis, I., Vlahavas, I., Effective voting of heterogeneous classifiers, *In Proceedings of the 15th European Conference on Machine Learning (ECML)*, LNAI 3201, 465-476, Springer, 2004.

113. Valarakos, A., Paliouras, G., Karkaletsis, K., Vouros G., Enhancing Ontological Knowledge through Ontology Population and Enrichment, *In Proceedings of the 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, LNAI 3257, 144-156, Springer, 2004.

114. Vilalta, R., Drissi, Y., A perspective view and survey of meta-learning, *Artificial Intelligence Review*, 18(2), 77-95, 2002.

115. Viterbi, A. J., Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, *IEEE Transactions on Information Theory*, 13, 260-269, 1967.

116. Vlahavas, I., Kefalas. P., Bassiliades, N., Refanidis, I., Kokkoras, F., Sakellariou, H., *Artificial Intelligence*, Gartaganis Publications, Thessaloniki, Greece, 2002.

117. Wang, Y., Witten, I.H., Induction of model trees for predicting continuous classes. *In Proceedings of the poster papers of the European Conference on Machine Learning (ECML)*. Un. of Economics, Faculty of Informatics and Statistics, Prague, 1997.

118. Witten, I.H., Bell, T.C., The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression, *IEEE Transactions on Information Theory*, 37(4), 1991.

119. Witten, I., Frank, E., Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, *Morgan Kaufmann Publishers*, 2000.

120. Wolpert, D., Stacked Generalization, *Neural Networks*,5(2): 241-260, 1992.

121. Wolpert, D., On the connection between in-sample testing and generalization error, *Complex Systems*, 6, 47-94, 1992b.

122. Yih, W-T., Template-based information extraction from tree-structured HTML documents. *Master's thesis*, National Taiwan University, 1997.

123. Ženko, B., Todorovski, L., Džeroski, S., A comparison of stacking with MDTs to bagging, boosting, and other stacking methods. *In Proceedings of the 1st IEEE International Conference on Data Mining (ICDM)*, 669-670, 2001.

ΓΛΩΣΣΑΡΙ

Αβεβαιότητα	Uncertainty
Ανάκληση	Recall
Αγνοούμενη τιμή	Missing value
Ακολουθιακά δεδομένα	Sequential data
Ακολουθιακή επικάλυψη	Sequential covering
Ακρίβεια	Precision
Αλγόριθμος αδύναμης μάθησης	Weak learner
Αναγνώριση σημασιολογικών ρόλων	Semantic role identification
Ανάκτηση πληροφορίας	Information retrieval
Ανάλυση αντιστοιχίας	Correspondence analysis
Ανομοιογένεια τιμών κλάσης	Class imbalance
Αριθμητική πρόβλεψη	Numeric prediction
Αρχιτεκτονική-διαιτητή	Arbiter-architecture
Αυτοδύναμη δειγματοληψία	Bootstrap sampling
Αυτόματο πεπερασμένης κατάστασης	Finite state automaton
Αυτό-μετάβαση	Self-transition
Βαθμοί εμπιστοσύνης	Confidence scores
Βαθμολόγηση ταξινομητών	Grading classifiers
Βασικό επίπεδο	Base-level
Γραμμική Παλινδρόμηση	Linear Regression
Δέντρα μετα απόφασης	Meta decision trees
Δέντρο Απόφασης	Decision Tree
Διαδικασία επαλήθευσης	Runtime
Διαδοχική γενίκευση	Cascade generalization
Διαστατικότητα	Dimensionality
Διασταυρωμένη Επικύρωση	Cross-Validation
Διαστρωμάτωση / Συστρωμάτωση	Stratification
Δικτυακός τόπος πωλητών	Vendor site
Δυαδική ταξινόμηση	Binary classification
Έγγραφο κειμένου	Text document
Εμπειρική πολυστρατηγική μάθηση	Empirical multistrategy learning
Εμπλουτισμός οντολογιών με παραδείγματα	Ontology Population
Ενδυνάμωση	Boosting
Έννοια στόχος	Target concept
Εντροπία	Entropy
Εμφωλίαση	Bagging
Εξαγωγή μονής θέσης πεδίου	Single-slot extraction
Εξαγωγή πολλαπλών θέσεων πεδίου	Multi-slot extraction
Εξόρυξη Γνώσης από Δεδομένα	Data Mining
Εξόρυξη Γνώσης από τον Πασκόσμιο Ιστό	Web Mining
Εξόρυξη Γνώσης από κείμενο	Text Mining
Εξυπηρετητής παγκοσμίου ιστού	Web server

Επαγωγική κλίση	Inductive bias
Επαγωγική μάθηση	Inductive learning
Επισημείωση	Annotation
Επισημείωση μερών του λόγου	Part-of-speech tagging
Ετερογενείς ταξινομητές	Heterogeneous classifiers
Ημι-δομημένος	Semi-structured
Θεματική περιοχή	Domain
Θέση στόχος	Target-slot
Καθαρισμός δεδομένων	Data cleaning
Κανόνες εισαγωγής ετικετών	Tagging rules
Κανόνες περιεχομένου	Contextual rules
Κανόνες συσχέτισης	Association rules
Κρυφά Μαρκοβιανά Μοντέλα	Hidden Markov Models
Λεκτική Μονάδα	Token
Μάθηση βασισμένη σε κανόνες	Rule-based learning
Μάθηση βασισμένη στα παραδείγματα	Instance-based learning
“Μαύρη Τέχνη”	“Black Art”
Μέση διαφορετικότητα	Average diversity
Μεσολαβητής πληροφορίας	Information mediator
Μετα-επίπεδο	Meta-level
Μετα-μάθηση	Meta-learning
Μηχανική Μάθηση	Machine Learning
Μηχανική Οντολογιών	Ontology Engineering
Μοντέλα δέντρων	Model trees
Μοντελοποίηση χρηστών	User modeling
Ομάδες συζητήσεων	Newsgroups
Ομογενείς ταξινομητές	Homogeneous classifiers
Οντολογία	Ontology
Οπτικοποίηση	Visualization
Όριο	Boundary
Παγκόσμιος Ιστός	World Wide Web
Παλινδρόμηση	Regression
Παράδειγμα	Instance
Πεδίο	Field
Περιεχόμενο κατανοητό από τη μηχανή	Machine readable content
Πιθανοτική Ψηφοφορία	Probabilistic Voting
Πίνακας ενδεχομένων	Contingency table
Πλειοψηφική Ψηφοφορία	Majority Voting
Πολύ-αποκριτικά μοντέλα δέντρων	Multi-response model trees
Πολύ-αποκριτική γραμμική παλινδρόμηση	Multi-response linear regression
Πολύ-στρατηγική μάθηση	Multistrategy learning
Πράκτορας σύγκρισης αγοράς	Shopping comparison agent
Πράκτορες	Agents
Σημασιολογικός ιστός	Semantic web

Σημασιολογικό πλαίσιο	Semantic frame
Σταθμισμένη πλειοψηφική ψηφοφορία	Weighted majority voting
Στατιστική σημαντικότητα	Statistical significance
Σύνολο Δεδομένων	Dataset
Συντήρηση *wrappers*	Wrapper maintenance
Συσσωρευμένη ακολουθιακή μάθηση	Stacked sequential learning
Συσσωρευμένη γενίκευση	Stacked generalization
Συσσωρευμένη παλινδρόμηση	Stacked regression
Συσσώρευση	Stacking
Συσσώρευση δύο επιπέδων	Bi-level stacking
Σχεδιότυπο	Template
Σχετική βελτίωση	Relative improvement
Ταξινόμηση	Classification
Ταξινόμηση πολλαπλών κλάσεων	Multi-class classification
Ταξινομητής	Classifier
Τεχνητή Νοημοσύνη	Artificial Intelligence
Τοπικά σταθμική παλινδρόμηση	Locally weighted regression
Τυχαία δειγματοληψία με επανατοποθέτηση	Random sampling with replacement
Υπερ-πληροφόρηση	Information overload
Υπερ-σύνδεσμος	Hyperlink
Υπομνηματισμός	Annotation
Ψηφοφορία	Voting
Ψηφοφορία με χρήση πιθανοτικών κατανομών	Voting with probability distributions

GLOSSARY

Agents	Πράκτορες
Annotation	Επισημείωση, Υπομνηματισμός
Arbiter-architecture	Αρχιτεκτονική-διαιτητή
Artificial Intelligence	Τεχνητή Νοημοσύνη
Association rules	Κανόνες συσχέτισης
Average diversity	Μέση διαφορετικότητα
Bagging	Εμφωλίαση
Base-level	Βασικό Επίπεδο
Bi-level stacking	Συσσώρευση δύο επιπέδων
Binary classification	Δυαδική ταξινόμηση
“Black Art”	“Μαύρη Τέχνη”
Boosting	Ενδυνάμωση
Bootstrap sampling	Αυτοδύναμη δειγματοληψία
Boundary	Όριο
Cascade generalization	Διαδοχική γενίκευση
Class imbalance	Ανομοιογένεια τιμών κλάσης
Classification	Ταξινόμηση
Classifier	Ταξινομητής
Confidence scores	Βαθμοί εμπιστοσύνης
Contextual rules	Κανόνες περιεχομένου
Contingency table	Πίνακας ενδεχομένων
Correspondence analysis	Ανάλυση αντιστοιχίας
Cross-Validation	Διασταυρωμένη Επικύρωση
Data cleaning	Καθαρισμός δεδομένων
Data Mining	Εξόρυξη Γνώσης από Δεδομένα
Dataset	Σύνολο Δεδομένων
Decision Tree	Δέντρο Απόφασης
Dimensionality	Διαστατικότητα
Domain	Θεματική περιοχή
Empirical multistrategy learning	Εμπειρική πολυστρατηγική μάθηση
Entropy	Εντροπία
Field	Πεδίο
Finite state automaton	Αυτόματο πεπερασμένης κατάστασης
Grading classifiers	Βαθμολόγηση ταξινομητών
Heterogeneous classifiers	Ετερογενείς ταξινομητές
Hidden Markov Models	Κρυφά Μαρκοβιανά Μοντέλα
Homogeneous classifiers	Ομογενείς ταξινομητές
Hyperlink	Υπερ-σύνδεσμος

Inductive bias	Επαγωγική κλίση
Inductive learning	Επαγωγική μάθηση
Information mediator	Μεσολαβητής πληροφορίας
Information overload	Υπερ-πληροφόρηση
Information retrieval	Ανάκτηση πληροφορίας
Instance	Παράδειγμα
Instance-based learning	Μάθηση βασισμένη στα παραδείγματα
Linear Regression	Γραμμική Παλινδρόμηση
Locally weighted regression	Τοπικά σταθμική παλινδρόμηση
Machine Learning	Μηχανική Μάθηση
Machine readable content	Περιεχόμενο κατανοητό από τη μηχανή
Majority Voting	Πλειοψηφική Ψηφοφορία
Meta decision trees	Δέντρα μετα απόφασης
Meta-learning	Μετα-μάθηση
Meta-level	Μετα-επίπεδο
Missing Value	Αγνοούμενη τιμή
Model trees	Μοντέλα δέντρων
Multi-class classification	Ταξινόμηση πολλαπλών κλάσεων
Multi-response linear regression	Πολύ-αποκριτική γραμμική παλινδρόμηση
Multi-response model trees	Πολύ-αποκριτικά μοντέλα δέντρων
Multi-slot extraction	Εξαγωγή πολλαπλών θέσεων πεδίου
Multistrategy learning	Πολύ-στρατηγική μάθηση
Newsgroups	Ομάδες συζητήσεων
Numeric prediction	Αριθμητική πρόβλεψη
Ontology	Οντολογία
Ontology Engineering	Μηχανική Οντολογιών
Ontology population	Εμπλουτισμός οντολογιών με παραδείγματα
Part-of-speech tagging	Επισημείωση μερών του λόγου
Precision	Ακρίβεια
Probabilistic Voting	Πιθανοτική Ψηφοφορία
Random sampling with replacement	Τυχαία δειγματοληψία με επανατοποθέτηση
Recall	Ανάκληση
Regression	Παλινδρόμηση
Relative improvement	Σχετική βελτίωση
Rule-based learning	Μάθηση βασισμένη σε κανόνες
Runtime	Διαδικασία επαλήθευσης
Self-transition	Αυτό-μετάβαση
Semantic frame	Σημασιολογικό πλαίσιο
Semantic role identification	Αναγνώριση σημασιολογικών ρόλων
Semantic web	Σημασιολογικός ιστός
Semi-structured	Ημιδομημένος

Sequential Covering	Ακολουθιακή επικάλυψη
Sequential data	Ακολουθιακά δεδομένα
Shopping comparison agent	Πράκτορας σύγκρισης αγοράς
Single-slot extraction	Εξαγωγή μονής θέσης πεδίου
Stacked generalization	Συσσωρευμένη γενίκευση
Stacked regression	Συσσωρευμένη παλινδρόμηση
Stacked sequential learning	Συσσωρευμένη ακολουθιακή μάθηση
Stacking	Συσσώρευση
Statistical significance	Στατιστική σημαντικότητα
Stratification	Διαστρωμάτωση / Συστρωμάτωση
Tagging rules	Κανόνες εισαγωγής ετικετών
Target Concept	Έννοια στόχος
Target-slot	Θέση στόχος
Template	Σχεδιότυπο
Text document	Έγγραφο κειμένου
Text Mining	Εξόρυξη Γνώσης από κείμενο
Token	Λεκτική Μονάδα
Uncertainty	Αβεβαιότητα
User modeling	Μοντελοποίηση χρηστών
Vendor site	Δικτυακός τόπος πωλητών
Visualization	Οπτικοποίηση
Voting	Ψηφοφορία
Voting with probability distributions	Ψηφοφορία με χρήση πιθανοτικών κατανομών
Weak learner	Αλγόριθμος αδύναμης μάθησης
Weighted majority voting	Σταθμισμένη πλειοψηφική ψηφοφορία
Web Mining	Εξόρυξη Γνώσης από τον Πασκόσμιο Ιστό
Web server	Εξυπηρετητής παγκοσμίου ιστού
World Wide Web	Παγκόσμιος Ιστός
Wrapper maintenance	Συντήρηση *wrappers*