The Less the Merrier: Dimensionality Reduction and Knowledge Discovery

Keywords: knowledge discovery, dimensionality reduction, data preprocessing

An increasing number of contemporary applications produce massive volumes of very high dimensional data. In scientific databases, for example, it is common to encounter large sets of observations, represented by hundreds or even thousands of variables. Typical cases include astronomical, energy and network applications. In order to extract knowledge from these datasets, we need to access the underlying, hidden information. However, the size and dimensionality of these collections makes their processing and analysis impractical or even ineffective. Therefore, scaling up knowledge discovery algorithms for data of both high dimensionality and cardinality has been recently recognized as one of the top-10 problems in data mining research.

Problems associated with high dimensional data processing stem from three important factors, namely the curse of dimensionality, the empty space phenomenon and computational resources consumption. The curse of dimensionality and the empty space phenomenon are the main reasons that data mining algorithms exhibit poor results in terms of quality as dimensions grow while the computational resources consumption directly affects their scalability.

- The *curse of dimensionality* refers to the fact that in the absence of simplifying assumptions, the sample size needed to estimate a function of several variables to a given degree of accuracy in order to get a reasonably low variance estimate grows exponentially with the number of variables. An intuitive example is provided when considering the case of a square in R^2. Obviously, we need exactly 2^2 points in order to capture precisely the underlying structure and the relations among its points. However, when considering the case of a cube in R^3, we need 2^3 points while the four-dimensional hypercube necessitates 2^4 points.

- The *empty space phenomenon* on the other hand is a term which captures the sparsity of high dimensional spaces. In such spaces, the minimum and maximum distance of a dataset tend to be equal, therefore all data points seem to be concentrated at equal distance around a specific location. This phenomenon is directly related to the way we measure distances and is essentially influenced by the fact that each coordinate is attributed equal importance.

- *Computation resources* comprise an overarching term that encapsulates all requirements (e.g. time, space, disk, network,…) posed by an algorithm. Given a dataset of high dimensionality and cardinality, its evaluation will be time consuming, especially in cases where the employed algorithm is based on a distance metric. However, a transformation that embeds data in a lower dimensional space while preserving distances enables faster acquisition of the same results.

A way to avoid all these issues is perform a reduction of the input dimensions. Dimensionality reduction projects data from the original high dimensional space to a new lower dimensional space while retaining useful data properties such as pairwise distances or other statistical properties (e.g. variance). In formal notation, dimensionality reduction is a function f () with domain R^n and co domain R^k that embeds data from a high n-dimensional space to a predefined low k-dimensional space with $k << n$. The key issue is therefore the definition of the corresponding mapping function.

The minimization of the number of variables that describe an object is a difficult task since it directly implies that an amount of information will be lost. However, the reduction in the number of dimensions is plausible because:

1. A significant number of variables have minimal contribution therefore can be safely discarded.
2. Many variables are inter-dependent, in the sense that the behaviour of one directly affects the other; consequently they can be substituted by a single, new variable that comprises their linear combination.

Reducing dimensions can have a serious effect on the result of numerous KDD tasks. An excellent experimental validation of the curse of dimensionality and the empty space phenomenon came from the evaluation of the classification ability of k-NN algorithm on the REUTERS text collection. Reducing the number of variables to 0.05% of the initial dimensions incurred an amelioration in the quality of the obtained results that ranged from 20% to almost 100%. Another motivating result was obtained when evaluating the clustering quality of k-Means on a $5*10^5$ points, $5*10^2$-dimensional dataset; projecting to 2%- 10% of the initial dimensions reduced the running time of k-means from 7 minutes to less than 10 seconds while producing the same results.

In conclusion, dimensionality reduction is a powerful tool that facilitates the quality enhancement of many knowledge discovery algorithms. However, it should be used with caution, since it is not a one-fit solution. Therefore, particular effort should be paid in understanding the domain of application prior to preprocessing data with a focus on minimizing their variables.