MDM 2011, June 6 – June 9, Lulea, Sweden

Deploying In-Network Data Analysis Techniques in Sensor Networks*

George Valkanas, Alexis Kotsifakos
Dimitrios Gunopulos
Dept. of Informatics & Telecommunications
University of Athens, Greece

Ixent Galpin, Alasdair J.G. Gray,
Alvaro A.A. Fernandes, Norman W. Paton
School of Computer Science
University of Manchester, United Kingdom

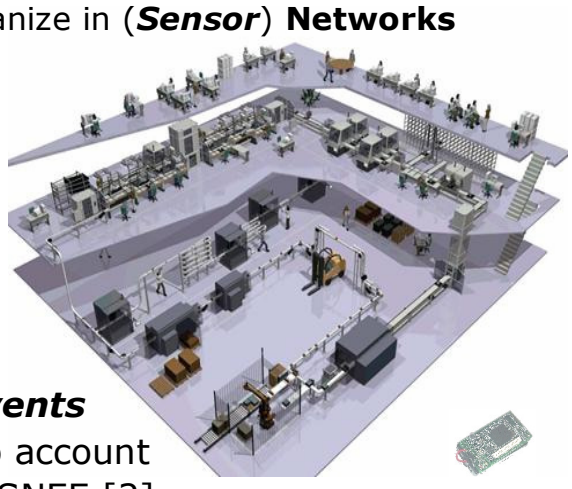
*Work has been supported by SemSorGrid4Env (FP7-223913) EU Project

Motivation

- **Sensors**
  - ✔ Can monitor inaccessible areas, high-performance infrastructures
  - ✔ Are used in numbers
  - ✔ Produce large amounts of data
  - ✔ Provide real-time readings
  - ✔ Communicate and self-organize in (***Sensor***) **Networks**
  - ✘ … but have limited power

Desiderata

- **In-Network Data Analysis**
  - Efficiently
  - Effectively
  - Real Time
- Intelligent data analysis, e.g. identify ***interesting events***
- Take battery limitations into account
- ***Ultimately*:** Integrate with SNEE [2]



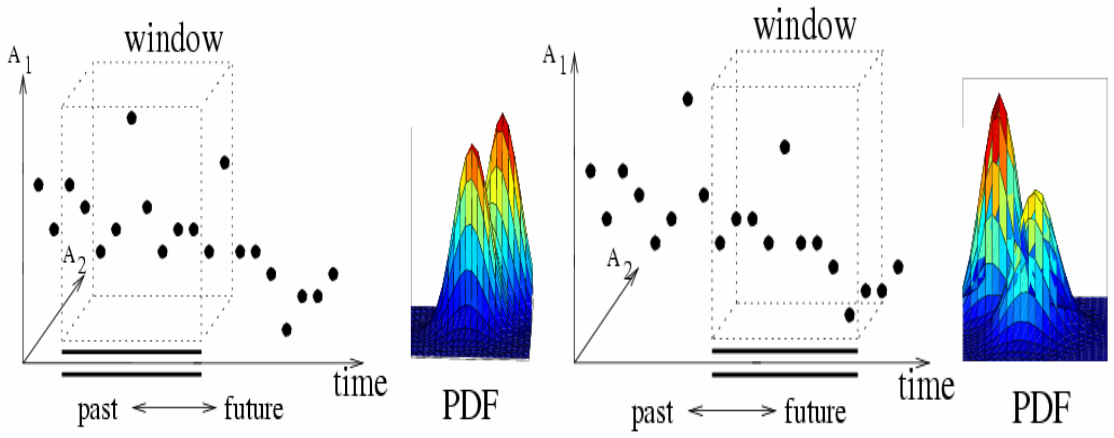
Our Demo

- Implemented two (2) ***Data Analysis Techniques***
  - Outlier Detection
  - Classification

Outlier Detection Background

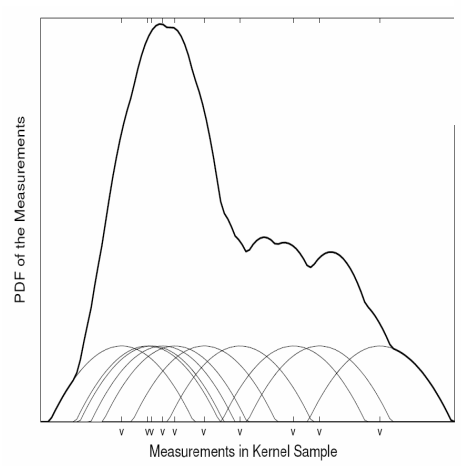
- **Intuition:** Detect abnormal behavior of sensed readings, i.e. ***outliers***
- **Definition:** Outliers are values that deviate significantly from the norm
- **Important** for:
  - Situation Detection (e.g. fire)
  - Focus on *interesting events only*
  - React to important readings -> Battery Savings!

- **Online Distributed Deviation Detection** (**D3**) [1]
  - Distance-based O(r,K) outlier discovery
  - Sliding Window model
  - Distributed processing
  - Online execution
  - Applicable in multi-dimensional data



Data Distribution Approximation

- Kernel Density Estimators
  - ✔ Generalization of Random Sampling
  - ✔ Effective approximation
  - ✔ Efficient online computation
  - ✔ Non-parametric
  - ✔ Adjusts to changes of input
  - ✔ Operates in a distributed fashion



D3 Kernel Density Estimator

- Epanechnikov Kernel
  - ✔ Closed form integral

$$\left(\frac{3}{4}\right)^d \frac{1}{B_1 \ldots B_d} \prod_{1 \le i \le d} \left(1 - \left(\frac{x_i}{B_i}\right)^2\right) \quad \text{if } \left|\frac{x_i}{B_i}\right| < 1$$

- Kernel Bandwidth **B** w/ **Scott's** rule

$$B_i = \sqrt{5}\sigma_i |R|^{-\frac{1}{d+4}}$$

D3 Algorithm

- Tuple **t** = sense the environment
- Sample on the input with ***chain sampling***
- Compute weight of tuple w/ *Epanechnikov*
- If *#neighbors* of **t** within radius **r** < **K**
  - Report **t** to parent as outlier

Classification Background

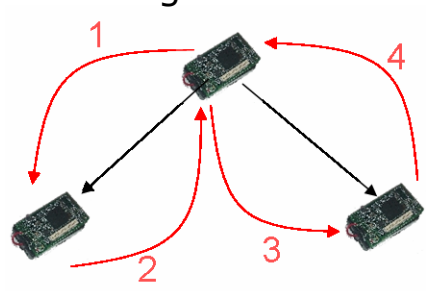
- Assume correlation of readings
- Important for
  - Missing value substitution
  - Communication reduction
  - Network Longevity

Linear Regression Classifier

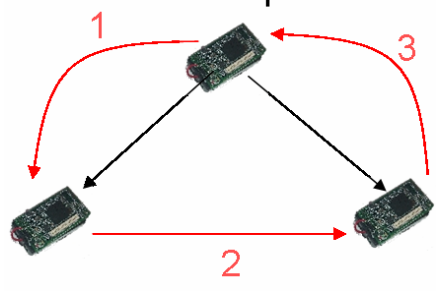
- Values are of the form: Y = ***a*** * X + ***b***
- Compute **(a, b)** efficiently

Implemented 3 communication protocols

1. Naïve DFS (everything to the sink node)
2. DFS with local computations
3. Sibling communication with local computations

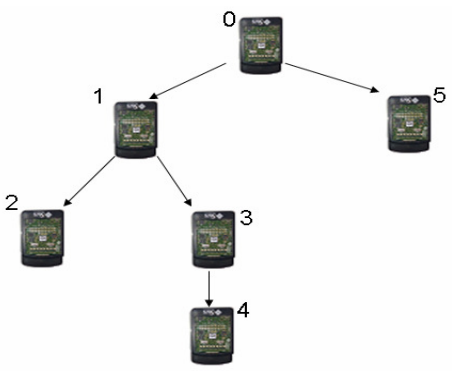


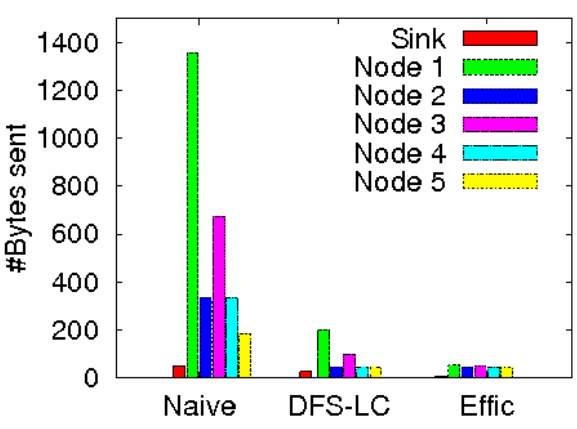
Cases 1 or 2



Case 3

Experimentation with SunSPOTs





References

[1] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos. Online outlier detection in sensor data using non-parametric models. VLDB'06

[2] Ixent Galpin, Christian Y. Brenninkmeijer, Alasdair J. Gray, Farhana Jabeen, Alvaro A. Fernandes, and Norman W. Paton. SNEE: a query processor for wireless sensor networks. Distrib. Parallel Databases, Feb. 2011