Outlier detection in Wireless Sensor Networks

By: Johan Becker & Fredrik Nilsson

Purpose of this presentation

Aims:

- Introduce Wireless Sensor Networks (WSN)
- Discuss challenges in WSNs
- Demonstrate with a case study how a distributed system for online detection of deviations can be constructed

Agenda

Introduction

Data acquisition

Data processing

Outlier detection

Case study: Online outlier detection in a distributed system of wireless sensors

Recap / Concluding remarks

Discussion

Introduction

Wireless Sensor Network

- What is it?
- Limitations

Main contribution:

- Provides a bridge between the physical and the digital world

Applications

- Meteorology (weather conditions)
- Monitor physical or environmental conditions
- Etc

Introduction

System model

- Central node called sink (or gateway)
  - Responsible for processing data
- Continuous data streams
  - The sensors are the source
- Data acquisition
  - How to collect the data from the sensors

Data Acquisition: Model-driven

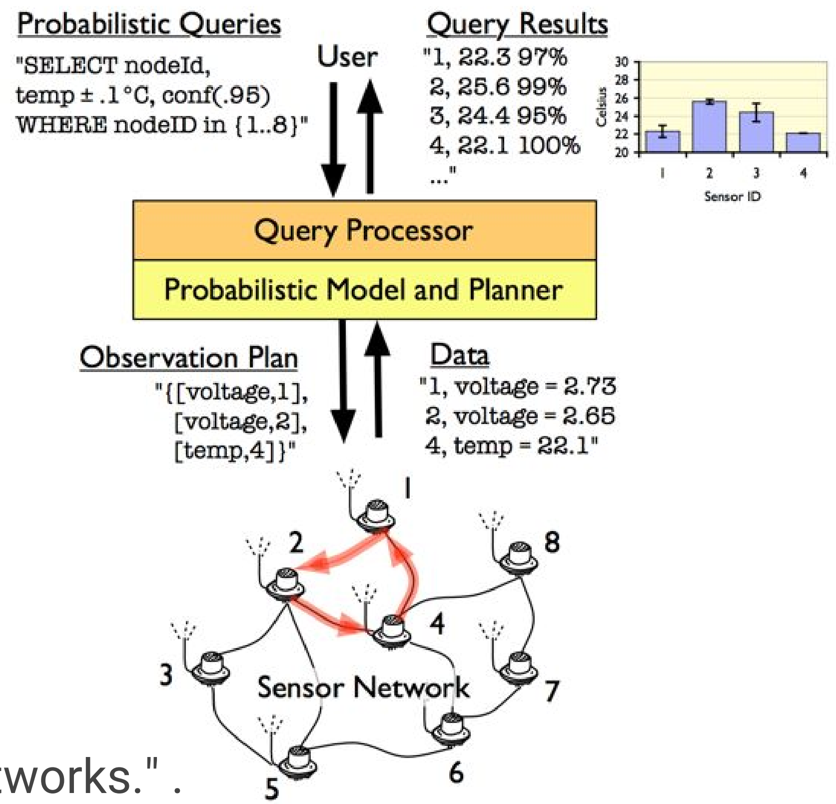
One statistical model in sink

Purpose:

Save energy by answering queries without contacting the sensors

First phase:

Train the statistic model in the sink



When queries arrive that demand better precision than what the model can provide, the model is updated

Image source:

Deshpande, Amol, et al.

"Model-driven data acquisition in sensor networks." .

Data Acquisition: Data-driven

Every sensor has a model for its readings

- Models are sent to sink

Whenever a sensor notices that its model is not accurate anymore, a new model is produced and sent to the sink

The sink is responsible for answering queries

- Energy savings and hard guarantees on errors
- Synchronized model

Data Acquisition: data series summarization

Smart way to compress data

- Amnesic functions

Most recent values must be accurate

- But historic values are allowed to be erroneous

Common in applications were nodes communicate sporadically

Example: Weather data

Data processing

What to do with the acquired data?

Two examples:

- Tracking of homogenous regions
- Detection of deviations/outliers

Focus will be on detection of outliers

Outliers

What is an outlier?

Why are outliers interesting?

Outlier detection: Approximate vs Exact approaches

What does approximate mean in this context?

Why not exact approaches?

3 classes of approximate approaches:

- Classification-based
- Node similarity-based
- Data distribution-based

Outlier detection: Classification-based

Bayesian classifiers to identify outliers.

Assumptions: a sensor’s current value is only influenced by it’s previous values and the current values of its closest neighbors.

Predict the expected range of next values

If the subsequent values are not within this range: deem as outliers.

Outlier detection: Node similarity-based

Two types of outliers:

Short simple outliers & long segmental outliers.

Identification can be performed by using the Discrete Wavelet Transform on the time series of the sensor values.

Compare original data with result of the transformation.

Certain threshold distance away: short outliers.

Long outliers

Data series are compared to the series from other nearby sensors.

Not within a given threshold distance then declare as a long outlier.

Outlier detection: Data distribution-based

Based on statistical properties and probability distributions.

- Sensors track their own distribution
- Use it to determine if a value is an outlier

Sensors can take advantage of other sensors' probability density functions

- To compare and identify outliers that are spatio-temporally correlated

Outlier detection: Data distribution-based

Two types of outliers

Distance-based outliers

A value is an outlier if it is further away from other values in the dataset given certain threshold.

Density-based outliers

Calculate Multi Granularity Deviation Factor and keep track of neighborhood, if value is significantly different report as outlier

Case study: Online identification of outliers in a distributed system of wireless sensors

Paper

Online outlier detection in sensor data using non-parametric models

By: S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos. 2006

Agenda:

- Recap
- Purpose
- Sensor network model / hierarchy
- Type of outliers
- Kernel estimation
- Approximation of distribution in a sliding window
- Detection of outliers
- Recap and Conclusion

Case study: recap

Resources are limited on a sensor

Approximation motivated by limited resources

- This is also the case for identification of outliers/deviations

Useful for finding broken sensors and anomalies in the network

- Highlights interesting events

Case study: Purpose

Aim: Identify outliers in real-time in a distributed fashion

How: Kernel density estimators to approximate the sensor data distribution
Calculate density of data-space around each value
Determine which values are outliers

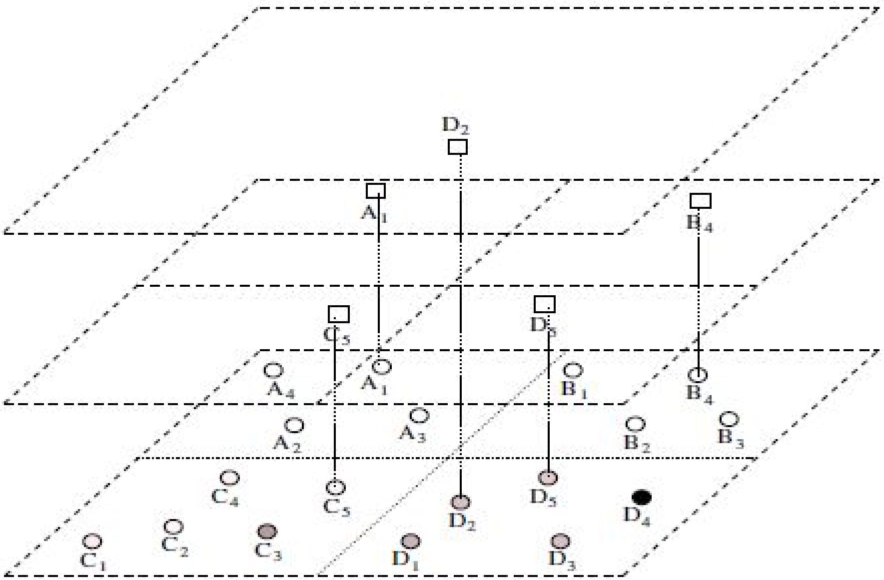
Outlier detection with:
Distance based algorithm
Local metrics based approach

Case study: Sensor network model

Hierarchy of the sensor network

Detect outliers at multiple levels

Fault tolerance



Case study: two types of outliers

Distance based outliers:

- Requires no prior knowledge of underlying data distribution.
- An outlier is an outlier if it is sufficiently far away from other values in the dataset.

Local metrics-based outliers:

- Calculate Multi Granularity Deviation Factor (MDEF)
- Keep track of neighborhood
- If value is significantly different report as outlier

Case study: Kernel estimation

Why kernel density estimators?

- Efficient to compute and maintain in a streaming environment
- Can effectively approximate an unknown data distribution
- Can easily be combined
- Scale well in multiple dimensions

What are kernel estimators?

Generalized form of sampling

Basic step is to produce a uniform random sample

Each point has weight of one distributed in the space around the sample point

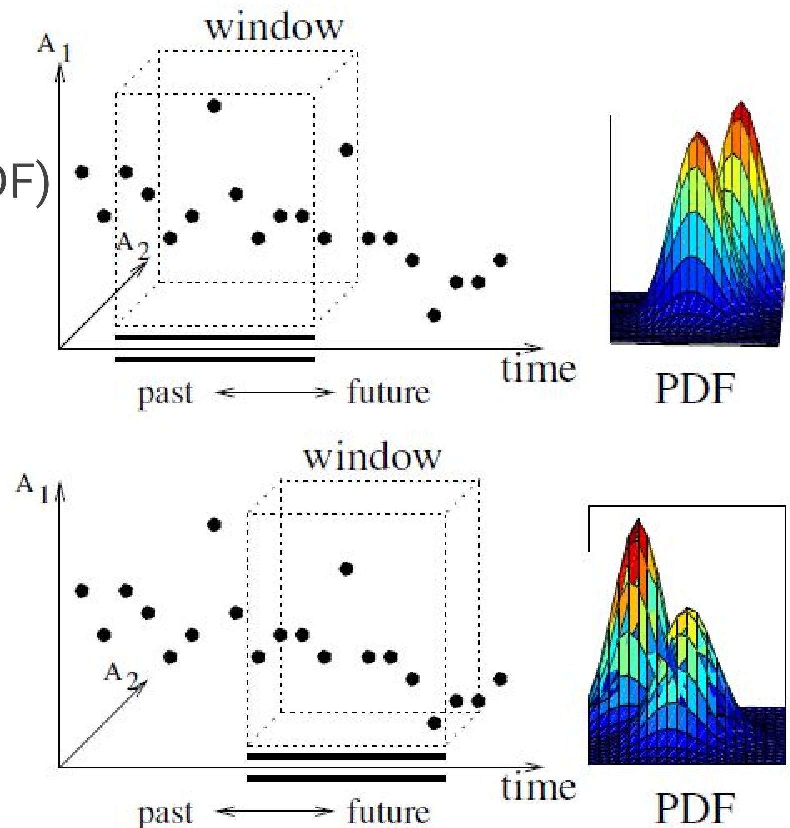
Case study: Distribution approximation in a Sliding window

Online approximation of the data distribution in a sliding window

Each sensor maintains a model of the distribution of values that it generates

- Done with kernel estimators and sampling, produces a Probability Density Function (PDF)

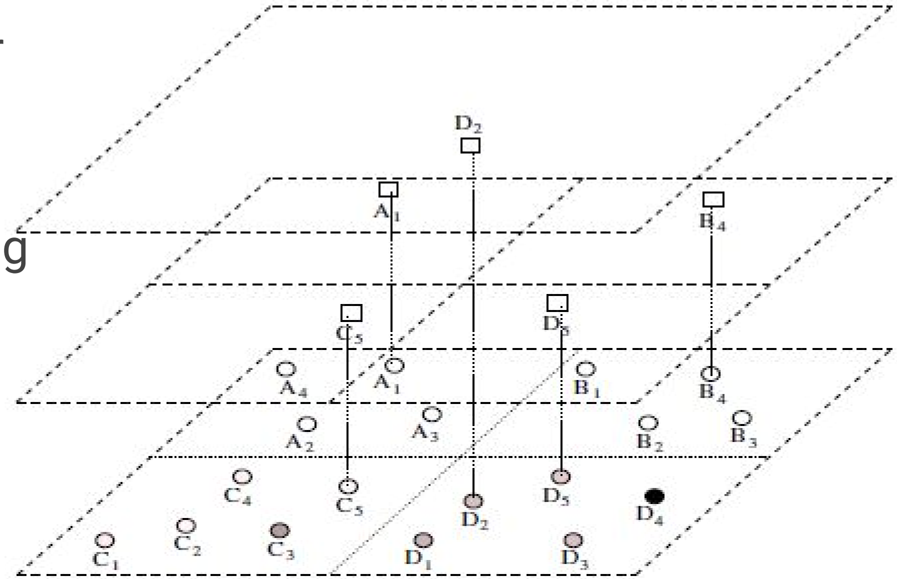
Every time the window moves a new PDF distribution is generated on a sample of the values that are currently inside the window.



Case study: Distributed detection of distance-based outliers

Detection of distance-based outliers
- Compare current value of a sensor with the probability density distribution function that it maintains
- If the surrounding density is not high enough, then the value is far away from the other values in the dataset
- If sufficiently far away from the other values, then flag it as an outlier

When a value has been flagged as an outlier
- Send the value to the parent
- The parent has a pool of outlier values
- Checks if outlier is an outlier considering the whole cell.
- If it is, send to parent's leader node



Case study: Outlier Detection using multi-granular local metrics

Detecting outliers using the multi-granular local metrics approach

- The surrounding neighbors’ values are considered already at the detection on the node.

How?

- A node has its own ***local*** *probability density function*, used to detect outliers
- When outlier is detected, compare with rest of system
- Comparison using a ***global*** *probability density function* stored on the node
- ***Global*** *probability density function* communicated via a leader node

The global probability function is constructed at the leader node

- Naive approach is that it sees all the values from the nodes in the cell
- Can be improved greatly

Case study: Recap / Conclusion

Aim: Identify outliers in real-time in a distributed fashion

How: Kernel density estimators to approximate the sensor data distribution
Calculate density of data-space around each value
Determine which values are outliers
Hierarchy

Outlier detection with:
Distance based algorithm
Local metrics based approach

Two types of outliers: Distance-based and local metrics-based

Recap / concluding remarks

Wireless sensor networks

Data acquisition

Data processing

Outlier detection

Case study: Online outlier detection in a distributed system of wireless sensors

Questions?

Questions?

Thank you for your attention!

Discussion