FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Influence of outliers in a railway remote monitoring system

Vítor A. Morais

DRAFT VERSION

Programa Doutoral em Engenharia Electrotécnica e de Computadores

Supervisor: António P. Martins

April 8, 2017

Influence of outliers in a railway remote monitoring system

Vítor A. Morais

Programa Doutoral em Engenharia Electrotécnica e de Computadores

April 8, 2017

Contents

CONTENTS

Symbols

kbps	Kilobit per second (often used kbit/s or kb/s) - bit rate
Mbps	Megabit per second (often used Mbit/s or Mb/s) - bit rate
Gbps	Gigabit per second (often used Gbit/s or Gb/s) - bit rate
dB	Decibel - Gain/Attenuation
kHz	Kilohertz - Frequency
MHz	Megahertz - Frequency
GHz	Gigahertz - Frequency
km	Kilometer - Distance
min	Minute - Time

Chapter 1

Introduction

This chapter presents the context, motivation and document structure of a study of outlier detection in a railways WSN-based smart grid.

1.1 Context and motivation

Smart grids are conceived as electric grids that deliver electricity from generation points to consumers, having the feature of controlling the entire process.

In railways...

Outliers are bla bla,.,.

The study of outliers is relevant due to it's influence in ....

With this work it is expected to raise the awareness of outliers detection in the phd study

1.2 Document structure

This document is divided in 4 chapters, each of them incorporate the relevant subsections to present the subjects mentioned

Table 1.1: Document structure

Chapter	Title
1	Introduction
2	Railways Remote Monitoring Systems
3	Outliers Detection
4	Future Research
5	Conclusions

Chapter 2

Railways Remote Monitoring Systems

In this chapter it is an overview of the railway system where the outliers detection is expected to be studied.

2.1 Smart Meters

2.2 Synthesis

Chapter 3

Outliers Detection

In this chapter it is made the study of the state of the art of outliers and it's relevance in railways.

3.1 Definition of outlier detection

Outlier detection is a computational task to detect and retrieve information from erroneous data values. The definition is usually close to anomaly detection or deviation detection.

3.2 Outlier detection in WSNs

Wireless sensor networks (WSNs) has been widely used in several applications in several domains such as industrial, scientific, medical and others. Those applications has been supported by the advances in wireless technologies as well as in the evolution of microcontroller technologies, with enhanced processing capabilities associated with reduced energy consumption.

3.2.1 Motivation

"In sensor networks, the majority of the energy is consumed in radio communication rather than computation" ... in the particular case of Sensoria sensors and Berkeley motes, the ratio of energy consumption between computation and communication modes is between 1000 and 10000 <rajasegarar2007>. Thus, an research opportunity is raised to reduce the communication usage of μCs by adding processing features towards the redution of energy consumption.

The motivation of detecting outliers in data acquired from WSNs has been extensivelly presented in the literature. The need for acquire data from harsh or "highly dynamic" environments as well as the need to validate and extract knowledge from the acquired data are one of the main points in the motivation to study the outlier detection in WSNs, **?**. <zang2010> <chandola2009> <ghorbel2015> <martins2015>

3.2.2 Research areas

Zhang et al. <zhang2010> identifies the outlier detection research areas in three domains:

- Intrusion detection: Situation caused by malicious attacks, where the detection techniques are query-driven techniques;
- Fault detection: Situation where the data suffer from noise and errors and where the detection techniques are data-driven ones;
- Event detection: Situation caused by the occurrence of one atomic or multiple events and where the majority of the research has been developed due to the complexity of detecting and extracting information on ?? upper layers ??

Based on the division of this three domains, the upcoming research is intended to be focused on the event detection techniques. ?? The railway environment requires closed subsystems that meets specific standards. Despite the intrusion detection should be considered, this must be took into consideration accordingly to the development and implementation of the wireless smart metering system for the railways. ?? The fault detection should and must be taken into consideration and the data outcome must be, preferably, a null value with a warning raised. ???

3.2.3 Challenges

The challenges of outlier detection in WSNs are related to the quality of the acquisition of the sensors, the fiability of the modules in terms of energy or environmental susceptibility, and the communication requirements and restrictions.

Zhang et al. <zhang2010> lists the challenges as the following:

- Resource constraints;
- High communication costs;
- Distributed streaming data;
- Dynamic network topology,
  frequent communication failures,
  mobility and heterogeneity of nodes;
- Large-scale deployment;
- Identifier outlier sources;

conclusion Copy paste from zhang:

Thus, the main challenge faced by outlier detection techniques for WSNs is to satisfy the mining accuracy requirements while maintaining the resource consumption of WSNs to a minimum [21]. In other words, the main question is how to process as much data as possible in a decentralized and online fashion while keeping the communication overhead, memory and computational cost low [1].

3.3 Classification of outlier

Zhang et al. <zhang2010> presents aspects to be used as metrics to compare characteristics of different outlier detection techniques. In parallel, Chandola et al. <chandola2009> presents a similar approach for the classification of outlier detection. In table 3.1 is present a comparison between two approaches to classify the nature of input sensor data.

Table 3.1: Classification of outlier techniques according to the nature of the input sensor data

<table>
<tr><th colspan="3">Zhang et al.</th><th colspan="3">Chandola et al.</th></tr>
<tr><td rowspan="5">Input sensor data</td><td rowspan="2">Attributes</td><td rowspan="2">univariate or multivariate</td><td rowspan="5">Nature of input data</td><td rowspan="2">Described using attributes</td><td>different types (binary, categorical, continuous)</td></tr>
<tr><td>quantity:
i) univariate;
ii) multivariate w/ same type;
iii) multivariate w/ different data types;</td></tr>
<tr><td rowspan="3">Correlations</td><td>dependencies among the attributes of sensor nodes</td><td rowspan="3">Related to each other</td><td>In sequence data, the data instances are linearly ordered, for example, time-series data, genome sequences, and protein sequences.</td></tr>
<tr><td rowspan="2">dependency of sensor node readings on history and neighboring node readings</td><td>In spatial data, each data instance is related to its neighboring instances, for example,vehicular traffic data, and ecological data.
When the spatial data has a temporal (sequential) component it is referred to as spatio-temporal data, for example, climate data.</td></tr>
<tr><td>In graph data, data instances are represented as vertices in a graph and are connected to other vertices with edges.</td></tr>
<tr><td rowspan="3"></td><td rowspan="3"></td><td rowspan="3"></td><td>Relationship</td><td>Can be categorized based on relationship present among data instances</td></tr>
<tr><td rowspan="2">Applicability</td><td>for statistical techniques</td></tr>
<tr><td>for nearest-neighbor-based techniques</td></tr>
</table>

Based on the work of Zhang et al. and Chandola et al., the table 3.2 identifies the different types of outliers. Those types differs on the objective of the outlier detection techniques: detect anomalies in individual data instances or in groups of data to detect irregularities, respectively, in local or in the global measuring system.

Table 3.2: Classification of the outlier techniques based on the type of the outlier/anomaly.

<table>
<tr><th colspan="3">Zhang et al.</th><th colspan="3">Chandola et al.</th></tr>
<tr><td rowspan="5">Type of outliers</td><td rowspan="2">Local outliers</td><td>Variation 1: anomalous values detection only depends on its historical values</td><td rowspan="5">Type of anomaly</td><td>Point anomalies</td><td>An individual data instance is considered anomalous, with respect to the others</td></tr>
<tr><td>Variation 2; anomalous values detection depends on historical values and on values of neighboring</td></tr>
<tr><td rowspan="3">Global outliers</td><td>Variation 1: All data is transmitted to a centralized architecture where outlier detection techniques takes place</td><td rowspan="3">Contextual anomalies</td><td>Contextual attributes: are used to determine the context for a given instance</td></tr>
<tr><td>Variation 2: Data from a cluster of sensors is used for outlier detection in a aggregate/clustering based architecture</td><td rowspan="2">Behavioral attributes: defines the noncontextual characteristics of a given instance.</td></tr>
<tr><td>Variation 3: Individual nodes can identify global outliers if they have a copy of global estimator model obtained from the sink node</td></tr>
<tr><td colspan="2"></td><td>Collective anomalies</td><td>If a collection of related data instances is anomalous with respect to the entire data set, it is defined as a collective anomaly.</td></tr>
</table>

Table 3.3 continues the classification, focusing in three parts:

- The need of pre-classified data (to implement supervised, semi-supervised or unsupervised outlier detection techniques);
- The output of outlier detection techniques (binary labels for normal/abnormal data-set and a score for each data-set to evaluate the weight of being an anomaly)
- The identity of the outliers (detect errors, events or malicious attacks)

Table 3.3: Classification of outlier detection techniques according to: i) need of pre-classified data; ii) output of detection techniques; iii) identity of outliers

<table>
<tr><th colspan="3">Zhang et al.</th><th colspan="3">Chandola et al.</th></tr>
<tr><td rowspan="3">Availability of pre-defined data</td><td>Supervised</td><td>Require pre-classified normal and abnormal data</td><td rowspan="3">Data labels: normal or anomalous</td><td>Labels obtained by Supervised Anomaly Detection</td><td>Training data has labeled instances for normal and anomalous classes</td></tr>
<tr><td>Semi-supervised</td><td>Require only pre-classified normal data</td><td>Labels obtained by Semi-supervised Anomaly Detection</td><td>Training data has labeled instances only for normal class. There is no labels for the anomalous classes</td></tr>
<tr><td>Unsupervised</td><td>Do not require pre-classified data</td><td>Labels obtained by Unsupervised Anomaly Detection</td><td>Techniques that do not require training data</td></tr>
<tr><td rowspan="2">Degree of being an outlier</td><td>Scalar</td><td>Zero-one classification: Classifies a data measurement into normal or outlier class</td><td rowspan="5">Output of Anomaly detection</td><td rowspan="2">Scores</td><td rowspan="2">Degree of which a data instance is consider an anomaly</td></tr>
<tr><td>Score</td><td>Assign to each data measurements a outlier score; Display a ranked list of outliers</td></tr>
<tr><td rowspan="3">Identity of outliers</td><td>Errors</td><td>Noise-related measurement or data coming from a faulty sensor</td><td rowspan="3">Labels</td><td rowspan="3">Provide binary labels (normal/anomalous)</td></tr>
<tr><td>Events</td><td>Particular phenomena that changes the real-world state</td></tr>
<tr><td>Malicious attacks</td><td>Outside of the scope (In the scope of network security)</td></tr>
</table>

3.4 Taxonomy of Outlier Detection Techniques

The study of detection techniques requires a well defined taxonomy framework that addresses the related work on the different areas. This taxonomy is well defined and solid in the literature, where the works of Zhang et al. and Chandola et al. reflect a similar approach on presenting a taxonomy for outlier detection techniques.

In the following sections a coverage in relevant techniques is presented:

- Classification based techniques.

  Bayesian Networks

  Rule-based techniques

  Support Vector Machines
- Statistical based techniques.

  Parametric — Gaussian based

  Non-parametric — Histogram based

  Non-parametric — Kernel function based
- Nearest Neighbor-based techniques.

  Using distance

  Using relative density
- Clustering based techniques.
- Spectral Decomposition based techniques.

3.5 Classification based techniques

Classification based techniques are based on systematic learning approaches based on sets of data. The supervised approaches requires knowledge to train a model (or classifier) from a set of data instances (or training data) and classifies a new data instance as normal or as outlier. The unsupervised approaches do not require knowledge and learn the boundary around normal instances, declaring the new instance as normal or as outlier depending if the data instance is outside of the boundary of the previous data sets.

The classification based techniques are listed as the following:

- Neural Networks-based;
- Bayesian Networks-based;
- Rule-based;
- Support Vector Machines-based.

Neural networks-based approaches are interesting strategies for outlier detection where a given neural network might be trained with only normal data-sets. At testing stage, the data instances that are similar to the training data-set are accepted by the neural network and then considered as normal. The remaining data-sets are rejected by the neural network due to their lack of similarity with normal data-sets. Thus, those data instances are considered as outliers. Based on the table 3.3, these techniques are classified as semi-supervised due to their need for normal data-sets for the training stage.

Bayesian networks-based approaches are identified as prominent techniques for outlier detection in WSNs, being the reason why they are extensively covered further on. Those techniques ...

Rule based ...

Support Vector Machine (SVM) relies on ...

3.5.1 Bayesian Networks

Zhang et al. <zhang2010> divide the bayesian network based techniques in three categories:

- Naïve Bayesian Networks;
- Bayesian Belief Networks;
- Dynamic Bayesian Network Models;

All those approaches uses probabilistic graphical models to represent a set of variables and their probabilistic interdependencies. This graphical model aggregates the information from different variables and provides an estimate on the expectancy of an event to belong to the learned class.

Xiang et al. <xiang2015> illustrates an application to measure the concentration of NO_2, CO and O_3 pollutants, using a bayesian network. All the three variables are all correlated and also

depends on the temperature as presented in figure 3.1. The real measurements acquired by the microcontroller are represented with (s) and the representations in (t) refers to the real concentration of those pollutants.

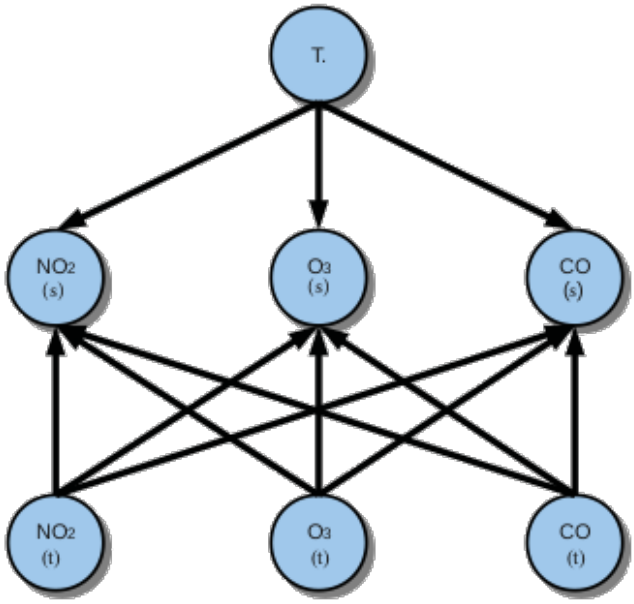

Figure 3.1: Application of a Bayesian Network to an atmospheric measurement system.

The three categories presented by Zhang et al. differs between them where the first category captures the sensor nodes correlations on spatio-temporal domain; The second one considers not only the spatio-temporal correlations but also the conditional dependence of sensor attributes; The third category proposes the measurement of state variables at a current time instance.

Janakiram et al. <janakiram2006> proposes the detection of outliers in sensor streamed data by capturing the conditional dependencies among the observation of it's attributes. this is made in three phases:

Training Phase Phase where the Bayesian Belief Network is trained to capture the spatio-temporal correlations.

Testing Phase Phase where the trained BBN is tested on the level of accuracy and, if needed, the learned parameters are updated.

Inference Phase Phase where the missing values are inferred and the remaining streamed data are tested to detect if it is an outlier or not.

Janakiram et al. also defined the BBN, where the BBN is a directed graph, together with an associated set of probabilistic tables. The graph is divided in nodes and arcs, where the nodes represents the variables and the arcs are the representation of the casual/influential relationship among variables.

The main contribution of BBN is the possibility to have a model that, with the dependency between uncertain variables (by filling a node probability table), it is possible to describe complex probabilistic reasoning about uncertainty.

Janakriam et al. describes their process in three steps:

- Constructing the Bayesian Belief Network

  IF a few variables have direct dependencies

  AND many of the variables are conditionaly independent

  THEN all the probabilities can be computed from the joint probability distribution.

- Learning Bayesian Belief Networks

  IF the network structure is given

  AND all variables are fully observable in the training examples

  THEN estimating the conditional probabilities is enough

  IF the network structure is given

  AND some of the variables are observable

  THEN apply neural network using Gradient Ascent Procedure

  IF the network structure is unknown

  THEN Use heuristic search

  OR Use constraint-based technique to search through potential structures

- Inferring from Bayesian Belief Networks

  THESIS The probability distribution of certain attributes might be inferred

  PROOF Given the fact that the values that other attributes can take are known

Paola et al <paola2014> proposes an adaptive distributed Bayesian approach for detecting outliers in data collected by a WSN. The focus of the proposed algorithm is the optimization of outlier classification accuracy, time and communication complexity and also considering externally imposed constraints on conflicting goals. The proposed algorithm is intended to run in each sensor node.

From the individual sensor node point of view, this algorithm consists in two phases:

Outlier detection Where, based on sensor readings and on the collaboration with neighbors, is made the probabilistic inference where the results are evaluated in three metrics: classification accuracy, time complexity and communication complexity.

Neighborhood selection Where the best neighbors are identified ad selected to cooperate with, and, in addiction, to correspond to a reconfiguration of the Bayesian Network structure.

In the global point of view, if there is a high number of cooperating nodes, the classification is natirally higher with the drawback if increasing the processing time and communication complexity (thus resulting in increased detection delay and increase of energy consumption).

Xiang et al. <xiang2016> proposes the addition of recover and recalibrate the drifted sensors simultaneously based on the usage of a Bayesian network.

The authors have applied their algorithm to the measurement of the variables in the sensor readings of the NO_2, CO and O_3 pollutants, as previously presented in figure 3.1. Based on the correlations of the sensor readings and on the temperature influence, the algorithm itself detects the outliers, recover valid information and adjust the BBN to automatically recalibrate the sensor.

3.5.2 Rule-based techniques

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

3.5.3 Support Vector Machines

3.6 Statistical based techniques

3.6.1 Parametric — Gaussian based

3.6.2 Non-parametric — Histogram based

3.6.3 Non-parametric — Kernel function based

3.7 Nearest Neighbor-based techniques

3.7.1 Using distance

3.7.2 Using relative density

3.8 Clustering based techniques

3.9 Spectral Decomposition based techniques

Chapter 4

Future Research

In this chapter there are presented the future steps in research on outliers detection on railways WSN-based smart grid.

4.1 Outliers detection definition

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

4.2 Synthesis

Chapter 5

Conclusion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Conclusion