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 15, 2017

Influence of outliers in a railway remote monitoring system

Vítor A. Morais

Programa Doutoral em Engenharia Electrotécnica e de Computadores

April 15, 2017

Contents

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 of PhD

The railway system is responsible for 1.3% of entire European energy consumption, Birol and Loubinoux (2016). The discussion of the energy efficiency in railways is a grown topic due to its contribution to the global energy consumption.

The energy efficiency analysis and management requires a detailed mapping of the energy consumption/generation in the railway system.

This detailed mapping of the energy flows should include, not only the rolling stock level but also the traction substations and the auxiliary services.

The knowledge of all the load curves permits the load prevision, peak shaving and energy cost optimization for all global railway system.

1.2 Shift2Rail Framework

This work is supported by the iRail PhD programme – Innovation in Railway Systems and Technologies whose objectives are aligned with the Shift2Rail objectives, Shift2Rail Joint Undertaking (2015):

- 1. Cutting the life-cycle cost of railway transport by as much as 50%;
- 2. Doubling the railway capacity;
- 3. Increasing the reliability and punctuality by as much as 50%.

Framed on the Shift2Rail (S2R) Innovation Programme 3 (IP3) with the focus on ”Cost efficient and reliable infrastructure”, it is proposed to develop a Smart Metering Demonstrator (SMD) that reach a detailed monitoring and supervision of various energy flows on the premises of embrace the entire Railway System.

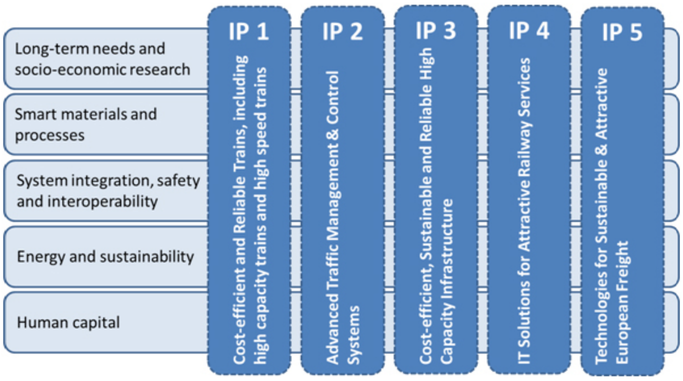

Figure 1.1: Shif2Rail Inovation Programmes.

The purpose of any energy management strategy is to build the dynamics of every loads and generators of the power system.

This should be performed based on an extensive knowledge of every energy flows.

This way, the SMD is required to propose and validate a standard metering architecture that involves the coordination of every measurements either in on-board and in ground. In advance, energy data analysis should be provided based on relevant stored data.

1.3 PhD state of the art

This section will cover a summary of the state of the art that supports this PhD.

Based on the state of the art, current metering systems focus on rolling stock on-board energy meters for energy billing purposes only, where the metering devices are located close to the pantograph, Shift2Rail Joint Undertaking (2015).

An advance beyond the state of the art is the expansion of the measurement system at railway system level, making it a distributed one, including both on-board and track-side measurements, thus achieving detailed mappings.

Other point in the state of the art is the intrusion level of currently used metering systems, that in one way, became a critical subsystem of the rolling stock and in other way, requires relatively long implementation, Shift2Rail Joint Undertaking (2015).

An advance beyond the state of the art is a solution based on non-intrusive technology. More detailed simulation models in conjunction with field measurements is the methodology to be investigated.

Specific challenges and requirements of this research are the development of non-intrusive Wireless Sensor Networks (WSN) in the railway environment. It is intended that this technology should be based on an open system and open interfaces for the data collection, aggregation and analysis. Issues like metering redundancy, outlier detection, fault tolerance and communication reliability, should be considered during the research. In addition, it is expected to design and specify a set of user applications. Those applications are focused in the energy analysis process with the aim of providing more information and detailed knowledge. It is expected that this detailed knowledge would be useful in a decision support system related with, in e.g., eco-driving strategies, timetable planning and preventive maintenance.

1.4 Influence of outliers in a railway remote monitoring system

Having in mind the state of the art that was previously presented in section 1.3, an important contribution of a wireless sensor network in the railway system is the availability of useful knowledge of the energy consumption to the decision support systems.

Therefore, such acquisition systems are required to provide accurate data regardless of the quality of the acquisition sensors, electromagnetic influences (EMI), sensor supply fluctuations, among others.

Through computational algorithms, the increasing of communication reliability and fault tolerance is possible. Those computational algorithms detects outliers or, in the scope of this PhD, detect erroneous data that will perturb the outcomes of decision support systems. Further on in chapter 2, this thematic is extensively explored.

1.5 Document structure

This document is divided in 5 chapters, each of them incorporate the relevant subsections to present the subjects mentioned.

Table 1.1: Document structure

Chapter	**Title**
1	Introduction
2	Outliers Detection
3	Future Research
4	Conclusions

Chapter 2

Outliers Detection

In this chapter it is made the study of the state of the art of outliers and it's relevance in railways. Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

2.1 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.

Branch et al. (2006) identifies the outlier detection as an essential step to either suppress or amplify outliers and precedes most any data analysis routine. Abid et al. (2016) points the need of detecting aberrant data and sensors within an WSN. Zhuang and Chen (2006) extends the outlier definition to the case where the outliers introduce in sensing queries and in sensing data analysis.

In the scope of the PhD and as previously presented in chapter 1, an outlier is a data value or a data instance that do not represent the correct consumption status.

The threshold of what is an outlier or not (or a value that do represent the correct consumption status or not) is given by the output of the subsystem that is immediately after the acquisition of consumption status subsystem, the decision support subsystem, gave a correct output or not. Figure 2.1 illustrates the integration of the consumption acquisition subsystems with the decision support subsystem.

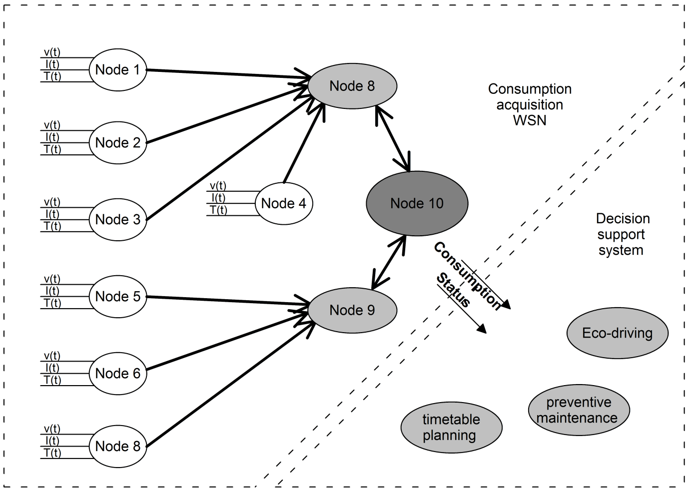

Figure 2.1: Integration of the WSN with a decision support system.

Without an outlier detection mechanism, the decision support subsystem may have the following outputs:

Input deviation from real value lower than a threshold The Decision Support Subsystem output is according to the real consumption conditions.

Input deviation from real value greater than a threshold The Decision Support Subsystem output is not according to the real consumption conditions.

The problem of taking decisions based on wrong considerations of the consumption status may lead to loss in desirable efficiency or increase of costs.

Let us consider a simple and hypothetical example where the DSS will provide an output towards suggesting an action in preventive maintenance based on the usage of a component. Considering that the usage of the component is depending on the counting of situations that the component is working above the nominal conditions. Without an outlier detection mechanism, the outliers will induce the DSS to count situations of overcharge of the component where the measurement is not related to the working above the nominal conditions but is related to external influences such as EMI or temperature. The output of DSS may suggest a preventive maintenance on a component that is working in proper conditions.

To conclude, with an outlier detection mechanism in the consumption acquisition subsystem the decision support subsystem may know if the value of consumption is an outlier or not and, with that information, the DSS output will be more accurate with the real conditions of operation.

2.2 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.

2.2.1 Motivation

Rajasegarar et al. (2007) points an important motivation for the inclusion of computational algorithms, i.e. outlier detection algorithms, to reduce the transmission of erroneous data, since in WSNs, the majority of the energy consumption occurs in the radio communication. In particular they present the case of Sensoria sensors and Berkeley motes where the energy consumption in communication exceeds in ranges from 1000 to 10000 the energy consumption of computation.

Thus, a research opportunity is raised to reduce the communication usage of μCs by adding processing features where the small increase in the computation will significantly reduce the energy consumption in the transmission. These processing features are, among others, the outlier detection algorithms.

On the field of the quality od the data acquired by WSNs, the motivation of detecting outliers in data acquired from WSNs has been extensively 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, Zhang et al. (2010); Chandola et al. (2009); Ghorbel et al. (2015); Martins et al. (2015).

2.2.2 Research areas

Zhang et al. Zhang et al. (2010) 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 it's complexity.

Based on the division of this three domains, the upcoming research is intended to be focused on the event detection and fault detection techniques. Specifically, the main goal for this research will be the event detection algorithms.

2.2.3 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. Zhang et al. (2010) 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;

Branch et al. (2006) identifies an important challenge, where the probability of occurrence of outlier events are extremely small. Abid et al. (2016) as well as **?** identifies the large amount of data as the main challenge for outlier detection in WSN. Zhuang and Chen (2006) identifies the inexpensive and low fidelity sensors as the main reason for the error generation and, the main challenge are identified on the distributed streaming data among a large amount of sensors. Ghorbel et al. (2015) points a main challenge as the processing of data from sensors that generates continuously data that is uncertain and unreliable.

To conclude, and in the scope of the PhD, the main challenges will be the usage of inexpensive and low fidelity sensors that will be affected by the rush railway environment. Complementary, the main challenge of using outlier detection mechanisms in the railway WSN is the balance between the detection accuracy and the influence that undetected data-instances will induce in other subsystems (in particular in decision support systems dependent on data from the WSN). In addition the detection accuracy is directly related with the memory usage, computational requirements, communication overhead, etc.

2.3 Classification of outlier

Zhang et al. (2010) presents aspects to be used as metrics to compare characteristics of different outlier detection techniques. In parallel, Chandola et al. (2009) presents a similar approach for the classification of outlier detection. In table 2.1 is present a comparison between two approaches to classify the nature of input sensor data.

Table 2.1: 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="8">Input sensor data</td><td rowspan="2">Attributes</td><td rowspan="2">univariate or multivariate</td><td rowspan="8">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="6">Correlations</td><td>dependencies among the attribures 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="5">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>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. (2010) and Chandola et al. (2009), the table 2.2 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 2.2: 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="6">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="6">Type of anomaly</td><td rowspan="2">Point anomalies</td><td rowspan="2">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 2.3 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 2.3: 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>

2.4 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. (2010) and Chandola et al. (2009) 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.

2.5 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 2.3, 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 ...

2.5.1 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 2.2. The real measurements acquired by the microcontroller are represented with (s) and the representations in (t) refers to the real concentration of those pollutants.

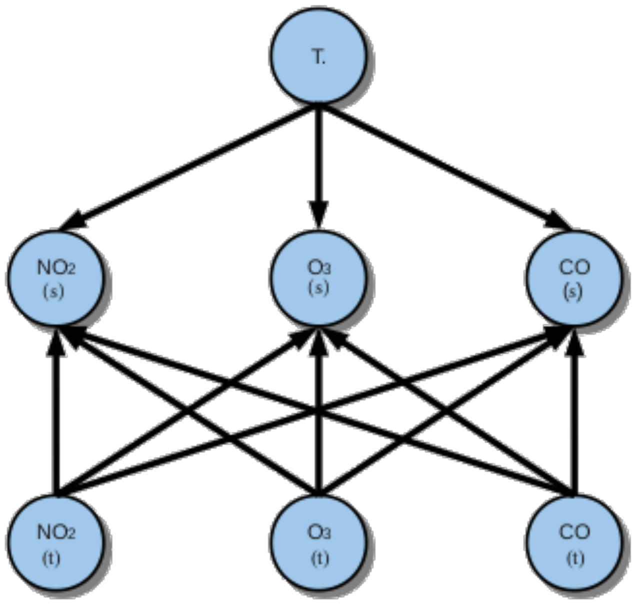

Figure 2.2: Application of a Bayesian Network to an atmospheric measurement system.

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 naturally 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 2.2. 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.

2.5.2 Rule-based techniques

Rule based is another classification based technique for outlier detection. Similarly, this technique is based on a training stage from a data-set and a model generation to detect new data-instances based on history values.

Rule based techniques depends on two steps <chandola2009>:

- **Learn rules from the training data-set**
  Using a learning algorithm (i.e. RIPPER, Decision Tree, etc.)
  Where each rule has an associated confidence value proportional to the ratio:
  $\text{Confidence Value} = \frac{\text{number of training instances correctly classified by the rule}}{\text{number of total training instances covered by the rule}}$
- **Find for each test instance the rule**
  That better capture the given test instance.
- **The anomaly score is**
  The inverse of the confidence value for the rule that better capture the test instance.

Islam et al. 2016 <islam2016> proposes an algorithm for outlier detection inserted in rule-based taxonomy. They propose a new belief-rule-based association rule, with the focus on handling various types of uncertainties.

Due to the nature of the sensor data, a traditional inference mechanism can not be used. Therefore they propose a new inference mechanism for the rule-based algorithm that consists of an input transaction databased that is converted into the following:

- belief transaction database;
- support calculation;
- belief matrix;
- confidence calculation;
- belief association rule discovery.

2.5.3 Support Vector Machines

Rather than performing outlier detection in the central node, Rajasegarar et al. <rajasegarar2007> proposes a distributed approach to:

- performs detection on local data at each node
- and communicates only the summary information to perform the global classification of the data.

Their proposal is based on a one-class quarter sphere svm and is divided into 2 parts:

- **Anomaly detection algorithm**

  The OD is supported by previous works where, with the fitness approach of a hypersphere to the data in a higher dimensional space, and by applying a linear optimization to the problem of fitting the hypersphere with minimal radius R, having the center fixed at the origin and encompassing the majority of the image vectors.

  The result of the linear optimization problem is the classification of the image vectors as:

  $\rightarrow$ **Support Vectors**, if inside the sphere;

  $\rightarrow$ **Outliers**, otherwise.
- **Distributed anomaly detection**

  1. Each sensor node runs the entire AD algorithm on local data;
  2. The resulting radius is sent to the parent node;
  3. Each parent computes the global radius;
  4. Parents sends the radius to children nodes;
  5. Children compares global radius with local one and updates parameters.

Xu et al. 2012 <xu2012> proposes a KNN-SVM which is a Support Vector Machine based on K-Nearest Neighbor Algorithm.

Despite KNN taxonomy is presented further on in section 3.7, in a synthesis the KNN is a distance-based approach that detect outliers in adata-instances lying in the sparsest regions or lying in the outside of a given model boundary of the feature space.

Considering the Quarter sphere SVM technique proposed by Rajasegarar et al. or by Sun et al 2008 <sun2008> the KNN-SVM combine the origin and the radius R that contain most of the samples and introduces kernel functions to make the optimization region more tighten.

Martins et al.<martins2015a, martins 2015b> has proposed a modifies SVM based on a kernel-based technique. In parallel, they propose an online sliding window scheme. The modification extends the original LS-SVM to be applied to the transient raw data collected from transmitters attached to a WSN. In a posterior work, Gil et al. <gil2016> compares the LS-SVM with PCA (a technique that will be presented further on in section 3.9).

2.6 Statistical based techniques

2.6.1 Parametric — Gaussian based

2.6.2 Non-parametric — Histogram based

2.6.3 Non-parametric — Kernel function based

2.7 Nearest Neighbor-based techniques

2.7.1 Using distance

2.7.2 Using relative density

2.8 Clustering based techniques

2.9 Spectral Decomposition based techniques

Chapter 3

Future Research

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

3.1 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.

3.2 Synthesis

Chapter 4

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.

Bibliography

Abid, A., A. Kachouri, and A. Mahfoudhi (2016, mar). Anomaly detection through outlier and neighborhood data in wireless sensor networks. In *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. Institute of Electrical and Electronics Engineers (IEEE).

Birol, F. and J.-P. Loubinoux (2016). 2016 edition of the uic-iea railway handbook on energy consumption and co2 emissions focuses on sustainability targets. Technical report, IEA - International Energy Agency; UIC - International Union of Railways.

Branch, J., B. Szymanski, C. Giannella, R. Wolff, and H. Kargupta (2006). In-network outlier detection in wireless sensor networks. In *26th IEEE International Conference on Distributed Computing Systems (ICDCS06)*. Institute of Electrical and Electronics Engineers (IEEE).

Chandola, V., A. Banerjee, and V. Kumar (2009, jul). Anomaly detection. *ACM Computing Surveys 41*(3), 1–58.

Ghorbel, O., W. Ayedi, H. Snoussi, and M. Abid (2015, jun). Fast and efficient outlier detection method in wireless sensor networks. *IEEE Sensors Journal 15*(6), 3403–3411.

Martins, H., L. Palma, A. Cardoso, and P. Gil (2015, may). A support vector machine based technique for online detection of outliers in transient time series. In *2015 10th Asian Control Conference (ASCC)*. Institute of Electrical and Electronics Engineers (IEEE).

Rajasegarar, S., C. Leckie, M. Palaniswami, and J. C. Bezdek (2007, jun). Quarter sphere based distributed anomaly detection in wireless sensor networks. In *2007 IEEE International Conference on Communications*. Institute of Electrical and Electronics Engineers (IEEE).

Shift2Rail Joint Undertaking (2015). Shift2rail joint undertaking multi-annual action plan. Technical report, Shift2Rail.

Zhang, Y., N. Meratnia, and P. Havinga (2010). Outlier detection techniques for wireless sensor networks: A survey. *IEEE Communications Surveys & Tutorials 12*(2), 159–170.

Zhuang, Y. and L. Chen (2006). In-network outlier cleaning for data collection in sensor networks. In *In CleanDB, Workshop in VLDB 2006*, pp. 41–48. APPENDIX.