International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161
©2014 INPRESSCO®, All Rights Reserved
Available at http://inpressco.com/category/ijcet

Research Article

An Outlier detection approach with data mining in wireless sensor network

M.Govindarajan[A] and V.Abinaya[A*]

[A]Department of computer science and engineering, Annamalai University, Chidambaram, India

Accepted 01 April 2014, Available online 15 April 2014, **Vol.4, No.2 (April 2014)**

Abstract

Wireless sensor networks had been deployed in the real world to collect large amounts of raw sensed data. However, the key challenge is to extract high level knowledge from such raw data. Sensor networks applications; outlier/anomaly detection has been paid more and more attention. The propose of a classification approach that provides outlier detection and data classification simultaneously. Experiments on Intel Berkley lab sensor dataset show that the proposed approach outperforms other techniques in both effectiveness & efficiency.

Keywords*: Outlier detection; data mining; wireless sensor network;*

1. Introduction

Outlier detection, also known as deviation detection or data cleansing, is a necessary pre-processing step in any data analysis application. Outlier detection in wireless sensor networks (WSNs) is the process of identifying those data instances that deviate from the rest of the data patterns based on a certain measure. The observations whose characteristics differ significantly from the normal profile are declared as outliers. Wireless sensor networks (WSNs) consists of hundreds or thousands of tiny, low-cost sensor nodes integrated with sensing, computational power, and short-range wireless communication capabilities, and have strong resource constraints in terms of energy, memory, computational capacity, and communication bandwidth. The large-scale and high density vision of the WSN implies that the network usually has to operate in a harsh and unattended environment. Moreover WSNs are vulnerable to faults and malicious attacks; this in turn causes inaccurate and unreliable sensor readings. Consequently, several factors make wireless sensor networks (WSNs) prone to outliers among those factors are:

- WSNs report the monitored data from the real world using imperfect sensing devices
- Such devices are battery powered and thus their performance tends to deplete as power is exhausted
- If WSN is deployed for military and security uses, sensors are exposed to manipulation by adversaries
- Since these networks may include a large number of sensors, this number may reach an extremely high value that can reach to million nodes depending on the application, hence the chance of error is more than that in traditional networks. An appropriate outlier detection technique for the WSN should pay attention to computing power, communication and storage limitations of the network and deal with the distributed nature of -data analysis.

A. Outlier detection

Outlier detection aims to find patterns in data that do not conform to expected behavior. It has extensive use in a wide variety of applications such as military surveillance for enemy activities, intrusion detection in cyber security, fraud detection for credit cards, insurance or health care and fault detection in safety critical systems. Their importance in data is due to the fact that they can translate into actionable information in a wide variety of applications.

B. Defining outliers

Outliers are patterns in data that do not conform to a well defined notion of normal behavior.

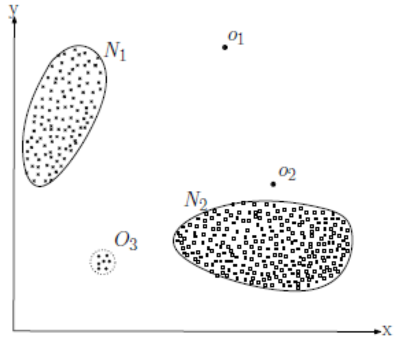

Figure 1. A simple example of ouliers in 2-dimentional dataset

Figure 1 illustrates outliers in a simple 2-dimensional data set. The data has two normal regions, N1 and N2, since

*Corresponding author **V.Abinaya** is a ME student most observations lie in these two regions. Points that are sufficiently far away from the regions, e.g., points o1 and o2, and points in region O3, are outliers. x y N1 N2 o1 o2 O3.

Outliers might be induced in the data for a variety of reasons, such as malicious activity, e.g., credit card fraud, cyber-intrusion, terrorist activity or breakdown of a system.

C. Issues

- Resource constraints
- High communication cost
- Distributed streaming data
- Dynamic network topology
- Large scale deployment
- Identifying outlier source

D. Applications

- Fraud detection- detecting fraudulent applications for credit cards, state benefits or detecting fraudulent usage of credit cards or mobile phones.
- Loan application processing - to detect fraudulent applications or potentially problematical customers.
- Intrusion detection- detecting unauthorised access in computer networks.
- Activity monitoring - detecting mobile phone fraud by monitoring phone activity or suspicious trades in the equity markets.
- Network performance- monitoring the performance of computer networks, for example to detect network bottlenecks.
- Fault diagnosis- monitoring processes to detect faults in motors, generators, pipelines or space instruments on space shuttles for\
- Structural defect detection- monitoring manufacturing lines to detect faulty production runs for example cracked beams.

2. Related work

A novel Intrusion Detection System (IDS) architecture utilizing both anomaly and misuse detection approaches. This hybrid Intrusion Detection System architecture consists of an anomaly detection module, a misuse detection module and a decision support system combining the results of these two detection modules. Simulation results of both anomaly and misuse detection modules based on the KDD 99 Data Set are given. It is observed that the proposed hybrid approach gives better performance over individual approaches (Ozgur Depren *et al*,2005). A Fraud can be seen in all insurance types including health insurance. Fraud in health insurance is done by intentional deception or misrepresentation for gaining some shabby benefit in the form of health expenditures (B. Melih Kirlidoga *et al* 2012). Intrusions pose a serious securing risk in a network environment. Network intrusion detection system aims to identify attacks or malicious activity in a network with a high detection rate while maintaining a low false alarm rate. Anomaly detection systems (ADS) monitor the behavior of a system and flag significant deviations from the normal activity as anomalies. In this paper the propose of anomaly detection method using “K-Means + C4.5”, a method to cascade k-Means clustering and the C4.5 decision tree methods for classifying anomalous and normal activities in a computer network(Amuthan Prabakar Muniyandia *et al*,2012). The comparisons of the outlier detection scoring measurements are based on the detection effectiveness. The results of the comparison prove that this approach of outlier detection is a promising approach to be utilized in different domain applications (Aiman Moyaid Said *et al*,2013). The propose of developing IDS for WSN have attracted much attention recently and thus, there are many publications proposing new IDS techniques or enhancement to the existing ones. This paper evaluates and compares the most prominent anomaly-based IDS systems for hierarchical WSNs and identifying their strengths and weaknesses (H.H. Soliman *et al*,2012).The k-Means clustering method is first used to partition the training instances into k clusters using Euclidean distance similarity. On each cluster, representing a density region of normal or anomaly instances and build decision trees using C4.5 decision tree algorithm. The decision tree on each cluster refines the decision boundaries by learning the subgroups within the cluster. To obtain a final conclusion of the results derived from the decision tree on each cluster (Bidyut *et al*,2012).Outlier detection is one of the main data mining tasks. The outliers in data are more significant and interesting than common ones in a wide variety of application domains, such as fraud detection, intrusion detection, ecosystem disturbances and many others (H. Jair Escalante *et al*,2004). Recently, a new trend for detecting the outlier by discovering frequent patterns (or frequent item sets) from the data set has been studied. In this paper present a summarization and comparative study of the available outlier detection scoring measurements which are based on the frequent patterns discovery (Hossein Moradi Koupaie *et al*,2013). Data mining tools and techniques can be used to detect fraud in large sets of insurance claim data. Based on a few cases that are known or suspected to be fraudulent, the anomaly detection technique calculates the likelihood or probability of each record to be fraudulent by analyzing the past insurance claims. The analysts can then have a closer investigation for the cases that have been marked by data mining software (Hamid Farvaresh *et al*,2011). An explosive growth in the field of wireless sensor networks (WSNs) has been achieved in the past few years. Due to its important wide range of applications especially military applications, environments monitoring, health care application, home automation, etc., they are exposed to security threats. Intrusion detection system (IDS) is one of the major and efficient defensive methods against attacks in WSN (R.Chetan *et al*,2012).

3. Existing methodology

The existing methodology was using the random tree classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of votes. In case of a regression, the classifier response is the average of the responses over all the trees in the forest .Get a prediction for each vector, which is oob relative to the last tree, using the very i-th tree. After all the trees have been trained, for each vector that has ever been oob, find the class-*winner* for it (the class that has got the majority of votes in the trees where the vector was oob) and compare it to the ground-truth response. Compute the classification error estimate as a ratio of the number of misclassified oob vectors to all the vectors in the original data. In case of regression, the oob-error is computed as the squared error for oob vectors difference divided by the total number of vectors.

4. Proposed methodology

The proposed methodology is used to detect the outliers and the detected outliers are classified as normal or outlier. It consists of four steps

Pre processing

Data pre-processing is an important step in the data mining process. Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis.

Outlier detection

Outlier detection is aim to detect the outlier based on following equation

$$\text{Detection Rate} = \frac{\text{Number of correctly classified instances}}{\text{total number of instances}} \text{X}100\% \qquad (1)$$

$$\text{False Alarm Rate} = \frac{\text{Number of incorrectly classified instances}}{\text{total number of instances}} \text{X}100\% \qquad (2)$$

Detection rate represented as number of normal data occur in the dataset and false alarm Rate is number of outlier data occur in the dataset.

Outlier classification

The decision tree classification is used to classify the sensor data, such as normal or outlier. The decision tree classification based on cross validation. The cross validation is sometime called rotation estimation. A model validation technique for accessing how the result of statistical analysis will generalize to an independent data set. Here $\frac{x-1}{x}$ data is used for training and $\frac{1}{x}$ is used for testing.

The data preprocessing is important step in the data mining. The data is collected incomplete, noisy and error. There is no quality of result; the quality is measure in term of accuracy. The detected outliers are classified as normal or outlier data.

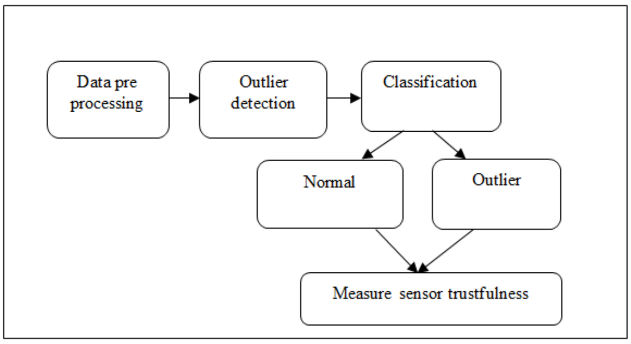

Figure 2. Block diagram of proposed work

Measure sensor trustfulness

Sensor trustfulness is based on classification accuracy, the classification accuracy is calculated from the dataset, here 699 instance are total number of instance, correctly classified instances are 617.the number of incorrectly classified instance are 82,such that the classification accuracy is 88%.sensor trustfulness are 88% evaluated on Intel Berkley lab dataset.

5. Experimental results

A. Description of the database

Table 1 Dataset schema

Data set schema		
Attributes of the dataset	**Date**	(yy-mm-dd)
	Time	(hh:mm:ss:xxxx)
	Epoch	(int)
	Moteid	(int)
	Temp	(real)
	Humidity	(real)
	Light	(real)
	Voltage	(real)

This dataset contains information about data collected from 54 sensors deployed in the Intel Berkeley Research lab between February 28th and April 5th, 2004. And also contain 699 instances, such instances are numeric and nominal values.

B. Performance measure

Detection Rate: It is defined as the ratio between the numbers of correctly classified instances to the total number of instances.

False Alarm Rate: It is defined as the ratio between the numbers of incorrectly classified instances to the total number of instances.

C. Before preprocessing

The Detection rate of before preprocessing is 76% and False alarm rate is 24%.

D. After preprocessing

The preprocessing filter is applied to the dataset the detection rate is increase to 88% and the false alarm rate is decrease to 12%.

E. Performance comparison

Table 2 Performance comparison

Performance comparison	**Performance comparison**		
	Preprocessing	**Detection rate**	**False alarm rate**
	Before	76%	24%
	After	88%	12%

The detection rate and False alarm rate are comparing between before and after preprocessing

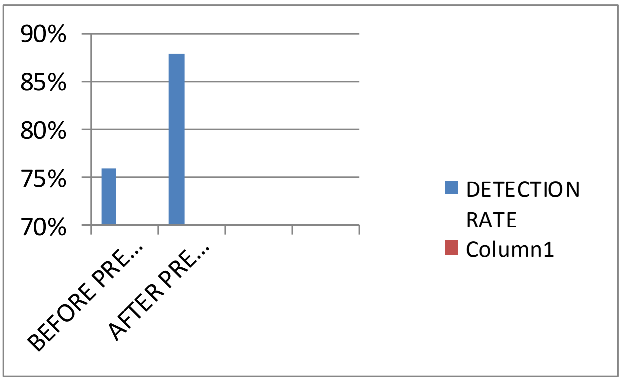

Figure3.detection rate comparison

Figure3 compare the detection rate between before and after preprocessing, the detection rate of before preprocessing is low and the detection rate of after preprocessing high.

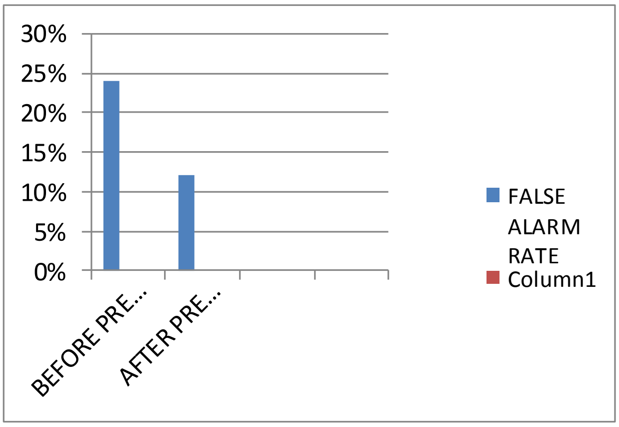

Figure4. False alarm rate comparison

Figure4 Compare the false alarm rate between the before and after preprocessing, the false alarm rate of before preprocessing is high and the false alarm rate of after preprocessing low.

Conclusion and future work

The wireless sensor data as outlier or normal based on decision tree based outlier detection technique and also classified the outlier data as normal data or error. The compared the performance of the technique with intelberkly research lab data, in this method finds the outlier with better accuracy for before preprocessing of 88%. In future find the solution for large dataset.

References

Ozgur Depren, Murat Topallar, Emin Anarim, M. Kemal Ciliz(2005),An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks ,*Expert System with applications*, 29,713-722.

Melih Kirlidoga,b, Cuneyt Asukb(2012), A fraud detection approach with data mining in health insurance, *Procedia - Social and Behavioral Sciences* ,62 , 989 – 994.

Amuthan Prabakar Muniyandia, R. Rajeswarib, R. Rajaramc(2012), Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm,international conference on communication Technology,*procedia engineering* ,30,174-182.

Aiman Moyaid Said, Dhanapal Durai Dominic and Brahim Belhaouari Samir(2013), Outlier Detection Scoring Measurements Based on Frequent Pattern Technique,*Research journel of applied sciences Engineering and Technology*,6(8),1340-1347.

H.H. Soliman a, Noha A. Hikal b, Nehal A. SakrA(2012), A comparative performance evaluation of intrusion detection techniques for hierarchical wireless sensor Networks,*Egyption informatics journal*,13,225-238.

Bidyut Kr. Patra(2012),Using the triangle inequality to accelerate Density based Outlier Detection Method, *Procedia Technology*, 6 , 469 – 474.

H. Jair Escalante (2004),A Comparison of Outlier Detection Algorithms for Machine Learning, *Computer Science Department National Institute of Astrophysics*, Optics and Electronics.

Hossein Moradi Koupaie, Suhaimi Ibrahim Javad Hosseinkhani(2013), Outlier Detection in Stream Data by Machine Learning and Feature Selection Methods, *International Journal of Advanced Computer Science and Information Technology (IJACSIT)*, 2, 17-24.

Hamid Farvaresh,Mohammad Mehdi Sepehri(2011),A datamining framework for detecting supscription fraud in telecommunication,*Engineering applications of artificial intelligence*,24,182-194.

Chetan R & Ashoka D.V(2012), Data mining based network intrusion detection system:A database centric approach,*International conference on computer communication and informatics*,10-12.