ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A Comparison of Outlier Detection Algorithm for Wireless Sensor Network

Mrs. Abinaya[*1], Mr. Govindarajan[2]

[*1]M.E final year in the Department of Computer Science&Engineering, Annamalai University, Chidambaram, India

[2]Assistant Professor, Department of Computer Science&Engineering Annamalai University, Chidambaram, India

abinayamvas@gmail.com

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,Decision tree.

Introduction

Outlier detection, also known as deviation detection or data cleansing, is a necessary preprocessing 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.

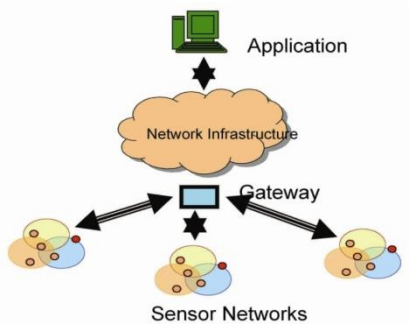

Figure1: Wireless sensor network

Consequently, traditional outlier detection techniques are not directly applicable to wireless sensor networks due to their particular requirements, dynamic nature, and resource limitations. 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. The main objective of outlier detection in WSNs thus is to identify outliers in the distributed streaming data in an online manner with high detection accuracy while maintaining the resource consumption of the network to a minimum Distinguishing between sources of outliers is a vital matter which in turn gives an insight on how to handle the detected outlier .Generally speaking if the detected outlier is an error or noisy data, it should be removed from the sensed data to ensure high data quality and accuracy; and to

ISSN: 2277-9655
Impact Factor: 1.852

save energy consumption by eliminating communication load.

Outlier Detection

Outlier detection is a critical task in many safety critical environments as the outlier indicates abnormal running conditions from which significant performance degradation may well result, such as an aircraft engine rotation defect or a flow problem in a pipeline.An outlier can denote an anomalous object in an image such as a land mine. An outlier may pinpoint an intruder inside a system with malicious intentions so rapid detection is essential. Outlier detection can detect a fault on a factory production line by constantly monitoring specific features of the products and comparing the real-time data with either the features of normal products or those for faults. It is imperative in tasks such as credit card usage monitoring or mobile phone monitoring to detect a sudden change in the usage pattern which may indicate fraudulent usage such as stolen card or stolen phone airtime. Outlier detection

Accomplishes this by analyzing and comparing the time series of usage statistics. For application processing, such as loan application processing or social security benefit payments, an outlier detection system can detect any anomalies in the application before approval or payment. Outlier detection can additionally monitor the circumstances of a benefit claimant over time to ensure the payment has not slipped into fraud. Equity or commodity traders can use outlier detection methods to monitor individual shares or markets and detect novel trends which may indicate buying or selling opportunities. A news delivery system can detect changing news stories and ensure the supplier is first with the breaking news. In a database, outliers may indicate fraudulent cases or they may just denote an error by the entry clerk or a misinterpretation of a missing value code, either way detection of the anomaly is vital for data base consistency and integrity.

Challenges of Outlier Detection

1. Modeling normal objects and outliers properly
   - Hard to enumerate all possible normal behaviors in an application
   - The border between normal and outlier objects is often a gray area
2. Application-specific outlier detection
   - Choice of distance measure among objects and the model of relationship among objects are often application-dependent
   - E.g., clinic data: a small deviation could be an outlier; while in marketing analysis, larger fluctuations
3. Handling noise in outlier detection
   - Noise may distort the normal objects and blur the distinction between normal objects and outliers. It may help hide outliers and reduce the effectiveness of outlier detection
4. Understandability
   - Understand why these are outliers: Justification of the detection

Issues

1. Resource constraints
2. High communication cost
3. Distributed streaming data
4. Dynamic network topology
5. Large scale deployment
6. Identifying outlier source

Applications

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

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 [1]. 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

ISSN: 2277-9655
Impact Factor: 1.852

the form of health expenditures[2]. 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[9]. 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[3]. 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 [6]. 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[7]. 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 [8]. 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 [4]. 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[10]. Therefore, 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 [5].

Existing Methodology

The existing methodology is used to detect and classify the outlier, the detected outlier are error in the sensor data. Sensor data are classified normal sensor data or error in the sensor data. Finally to measure the sensor trustfulness in terms of classification accuracy.

Random tree Algorithm

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 i-th 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.

Random Forest Algorithm

Random forests are an ensemble learning method for classification that operate by constructing a multitudeof decision tree at training
Time and outputting the class that is the mode of the classes output by individual trees. The shape of the decision to use in each node. The type of predictor to use in each leaf, the splitting objective to optimize in each node. The method for injecting randomness into the trees.
.

Disadvantages

1. The classification accuracy is low
2. The error in the sensor data is very high
3. High computational complexity
4. Execution speed is very high
5. Minimum memory usage

ISSN: 2277-9655
Impact Factor: 1.852

Proposed Methodology

The proposed methodology is used to detect the outliersand the detected outliers are classified as error or normal .it consists of four steps

Data 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

$$DetectionRate = \frac{Numberofcorrectly\ classifiedinstances}{totalnumberof\ instances} X100\%$$

$$FalseAlarmRate = \frac{Numberofincorrectly\ classifiedinstances}{totalnumberof\ instances} X100\%$$

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

Classification

The decision tree classification(random forest and random tree) is used to classify the sensor data, such as normal or outlier. The decision tree classification measuring sensor trustfulness 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.

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 619.the number of incorrectly classified instance are 80,so that the classification accuracy is 89%.sensor trustfulness are 89% evaluated on Intel Berkley lab dataset.

Block Diagram of Proposed Work

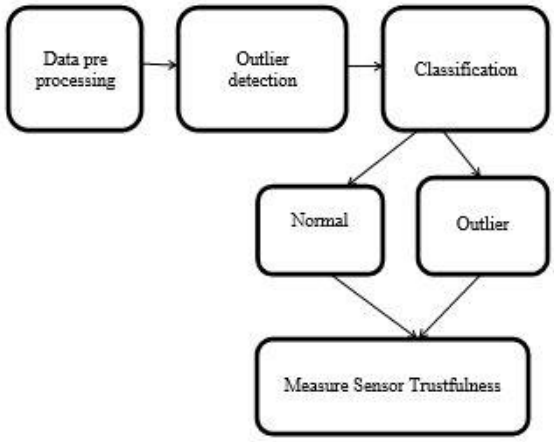

Figure 2: Block diagram of proposed work

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. Finally sensor trustfulness was calculated based on classification accuracy.

Advantages

1. Improve the classification accuracy
2. Minimize the error incorrectly classified instance
3. Improve the correctly classified instance

Experimental Results and Discussion

Description of the dataset

Table1 dataset Description

Date	Time	Epoch	Moteid	Temp	Humidity	light	voltage
(yy-mm-dd)	(hh:mm:ss.xxx)	(int)	(int)	(real)	(real)	(real)	(real)

ISSN: 2277-9655
Impact Factor: 1.852

This dataset contains information about data collected from 54 sensors deployed in the Intel Berkeley Research lab between February 28th and April 5th, 2004. This dataset was collected with epoch duration of about 30 s (resulting in a total of about 65,000 epochs) and it contains about 2.3 million readings.

Performance Measures

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

Before preprocessing of Random Tree

Before preprocessing the detection rate of random tree is 73% and the false alarm rate of random tree is 26%.

After preprocessing of Random Tree

After preprocessing the detection rate of random tree is 89% and the false alarm rate is 11%.The after preprocessing the detection rate was increased and the false alarm rate was decreased.

Before preprocessing of Random Forest

Before preprocessing the detection rate of random forest tree is 72% and the false alarm rate is 28%

After preprocessing of Random Forest

After preprocessing the detection rate of random forest tree is 81% and the false alarm rate is 19%. After preprocessing the detection rate was increased to 81% and the false alarm rate was decreased to 19%.

Performance Comparison

Table2 performance Comparison

Classifiers	Before pre processing		After pre processing	
	Detection rate	False alarm rate	Detection rate	False alarm rate
Random tree	**73%**	**26%**	**89%**	**11%**
Random forest	**72%**	**28%**	**81%**	**19%**

Before preprocessing both classifiers detection rate was low and false alarm rate was high, afterpreprocessing random tree and random forest tree of detection rate was increased and false alarm rate was decreased. Such that detection rate of random tree is 89% and the detection rate of random forest tree is 81%. So that the random tree detection rate and false alarm rate was better than the random forest classifier.

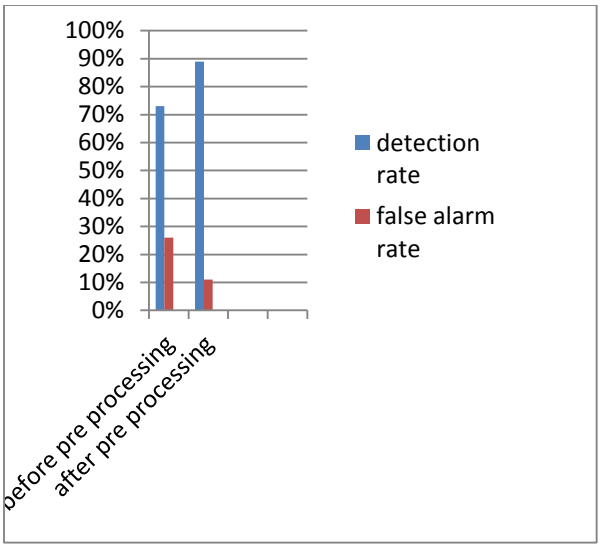

Figure 3:Random tree Performance Measures

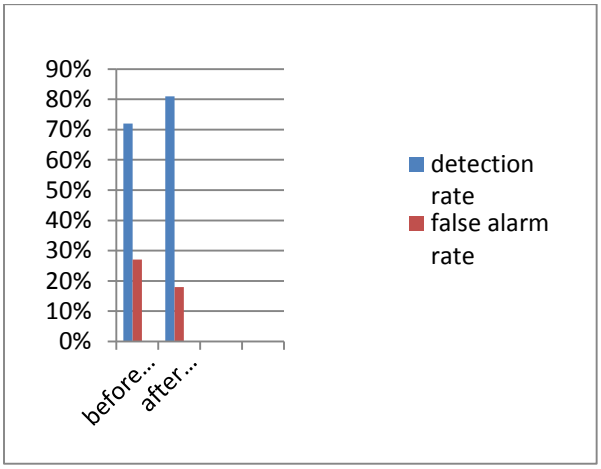

Figure 4:Random Forest Performance Measures

Conclusios

The wireless sensor data as outlier or normal based on decision tree based outlier detection technique and also classified the outlier data or normal data or. The compared the performance of the technique with Intel Berkley research lab data, in the random tree method finds the outlier with better accuracy is 89% and false alarm rate is 11%. The comparison between random tree and random forest tree, the random tree is best. In future, find the solution for large dataset.

ISSN: 2277-9655
Impact Factor: 1.852

References

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

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

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

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

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

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

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

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

[9] *Hamid Farvaresh,Mohammad Mehdi Sepehri "A datamining framework for detecting supscription fraud in telecommunication",engineering applications of artificial intelligence,24,2011,182-194.*

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