ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

A Review on Outlier Detection for Data Cleaning in Data Mining

Parmeet Kaur[1], Kanwarpreet Kaur

Student, Department of Computer Science, PTUKPIT Kapurthala, Punjab, India[1]

Student, Department of Computer Science, PU Regional Centre for IT and management Mohali, Punjab, India [2]

ABSTRACT: Data mining is the most important part of KDD (Knowledge Discovery in Database) process to find the peaceful information from huge of data. The main problem of data mining is error or misdate. So this problem can solve of data mining by using data cleaning. Data cleaning means is a process used to find out imprecise, imperfect or unreasonable data and then improve the quality through correcting of detected errors. Generally data cleaning reduces errors and improves the data quality. With the help of outlier detection we can detect the outlier and improve the quality of data .outlier detection is used to detect and remove the outlier from data. An object that does not follow the behavior of normal data object is called outliers. Outlier detection also plays an important role in data cleaning. Outlier detection is used in different applications like fraud detection, intrusion detection, track environmental changes, medical diagnosis so there is need to detect outliers from data. Various outlier detection techniques are used for data cleaning. Some of them use clustering based outlier detection approaches for data cleaning which can detect and remove can outliers from data. The Purpose of this paper is to review of different clustering approaches of outlier detection which is used for data cleaning.

KEYWORDS*:* data mining; data cleaning; clustering algorithm; outlier detection Algorithm.

I. INTRODUCTION

Nowadays large amount of data is generated from the various applications as number of user is increasing day by day. This Data is generated and saved in database which is ever-increasing at fast rate due to new technology and software, hardware improvements. For example humans generate different types of data like numeric data, text data, time and date data and also from documents, pictures, songs, movies, technical data and many other data into database. There is have to discover important data as useful patterns, association, relationships among these information in light of the fact that these extensive database may contain both valuable information and non-valuable.

Data mining[1] is the process of mining meaningful information, discovering new patterns, identifying relationship among data etc. from databases, Dataware house, graph data, flat files, relational database, spatial database, data repositories, normal data, network data, temporal database, data stream, business data, transactional database and World Wide Web data.

Outliers [2] are data objects which show different behavior than estimated behavior or same behavior as other data objects. Outliers data objects are the data objects which is different from the normal or ordinary objects or which indicates significant diversion from other data objects.. They are produced because of various reasons like noxious movement in system, instrumental mistake, natural changes, and blunders by human. Outliers are divided into 3 categories. [3]

Type 1, local outliers are the data objects which are different and isolated individual data objects with respect to all other data objects in data set. It is very easy to detect type 1 outlier data. Type 2, context outlier are the data objects which is segregated from other data objects in the same context. Semantic relationship among data points is referred by context of data objects. The basic Difference between type 1 outliers and type 2 outliers is that type 1 outlier is segregated from the entire other data objects in dataset of data objects rather than same context. Type 3 outliers are a subgroup or subset of group of data objects which come out as outliers with respect to entire dataset.[4]

There are various algorithms which are used for clustering data sets. From these algorithm K-Means algorithm is well known due to its simplicity and efficiency. This review paper is organized as follows: Application

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

areas of outlier detection are given in section II. Section III contains various approaches of outlier detection and conclusion is given in section V.

II. APPLICATION AREA OF OUTLIER DETECTION FOR DATA CLEANING

Outliers are generated from various applications area are due to some reasons. Outlier detection are used in different application for identify abnormal condition; remove noise, insurance clam, image processing medical application etc. the different application areas of outlier detection are given below.

Intrusion Detection: It identifies all abnormal patterns or malicious activity which indicates network or system attack or unusual behavior. Different types of intrusion detection are based on network, Host, stack. [5]

Aspect	Host Based instruction detection system	Network Based instruction detection system
Outliers In	OS Calls	Network Data.
Translates to	Malicious Code Unusual Behaviour Policy Violations	Denial of Network Services
Nature of Data analysis	Sequential	Point, Sequential, Collective
Granularity/ Profiling	User / Program	Packet Level/ NetFlows

Technique Used
Host Based Intrusion Detection Systems
Mixture of Models
Neural Networks[10]
Rule Based Systems
Network Based Intrusion Detection System
Statistical Profiling using Histograms
Parametric Statistical Modeling
Non-parametric Statistical Modeling
Bayesian Networks[11]
Support Vector Machines, Rule Based Systems
Neural Networks

Insurance clam outlier detection: it identifies the entire document which is submitted by the claimants. Neural network is used to detect these types of outliers.

Text data detection: database contains lots of data and some of them have noisy data. Such noisy data is removed by outlier detection technique from that data. Statistical Profiling using Histograms, neural network, clustering based, and support vector machine based outlier detection method is used to remove noise data.

Technique Used
Statistical Profiling using Histograms
Mixture of Models
Neural Networks
Support Vector Machines
Clustering Based

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Image processing outlier detection: Here aims of outlier detection are to detect changes in an image over time (motion detection) or in regions which appear abnormal on the static image. Satellite imagery, digit recognition, spectroscopy, mammographic image, and video surveillance are used to detect the outlier from image data.

Application Domain
Satellite Imagery
Digit Recognition
Mammographic Image Analysis
Spectroscopy
Video Surveillance

Medical Applications: A patient who have some diseases, there are some test like blood test, MRI scan, PET scan are conducted. Test data are collected from tests and which shows some abnormal data and also identify types of dieses and other related information. This is detected by medical diagnosis outlier detection methods.

Technique Used for fault detection in mechanical unit
Parametric Statistical Modelling
Non-Parametric Statistical Modelling
Neural Networks
Spectral
Rule Based Systems[12]

III. BASIC APPROACH TO DETECT THE OUTLIER DETECTION

There is several of clustering based approach to detect the outlier for cleaning. But some of mostly used are discussed below:

A. KMEANS CLSUTERING APPROACH TO DETECT OUTLIER

K-mean[8] is the simple and most efficient algorithm for outlier detection. kmeans describes that given dataset of n object divide into k cluster where K is desired number of cluster. An object o in one cluster is similar to objects belong in the same cluster and is called as intracluster similarity. An object o of one cluster is dissimilar with the objects of other cluster called as intercluster similarity. A centriod is defined for each cluster in kmean then data object are placed in cluster according minimum distance from centroids After processing all data object then kmean centriod is calculated again and again. For each iteration centriod of cluster changed according to their location. This process continues step by step until no centriod move.

Some demerits of kmeans is

1. In advance, need to specify k number of cluster
2. It is unable handle the noise or outlier or handle the cluster of different shapes

The complexity of k-mean clustering is O (IKN) where I denote number of iteration and k<<n[6]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Algorithm K-Means

Input parameter
D: data set contain n object
k: require number of cluster
Output parameter
k clusters containing n objects.
Algorithm
Step 1: Randomly select object k from dataset D as initial center of cluster
Step 2: for each data point in dataset do
Step 3: Calculate distance of data point from each cluster center
Step 4: Based on that distance find closest cluster and put that data point in that cluster
Step 5: end for
Step 6: After assigning the objects recalculate the mean of each clusters and update its value
Step 7: Go to
Step 2 until stopping criterion is match

B. KMEDIOD TO DETECT OUTLIER

Kmediod[9] is developed by Kaufman and Rousseuw in 1987. The algorithm chooses kmediod initially and then swaps the mediod object with non mediod as a result quality of cluster is improved. It is very robust when compare with kmean in the presence of noise or outlier. Algorithm work well with small dataset but does not work well with large dataset. The computational complexity of PAM is O (IK(NK)2) where I is a number of iteration[7]

Procedure of KMedioid

1. Input dataset d
2. Randomly select K object from D dataset.
3. Calculate total cost T for each pair of selected Si and non selected sh.
4. For each pair if TS I <_ 0 then it is replaced by SK
5. Then find similar mediod for each non selected object

Repeat the step 2, 3, 4 until find the mediod.

IV. CONCLUSION

There are various clustering algorithm to detect the outlier but from all the algorithm kmean is best algorithm rather than others because it is simple and efficient.with the help of kmeans we can detect the outlier and then remove that outlier for cleaning the data.

REFERENCES

1. Kaur, Parmeet, and Parmjeet Kaur. "AN OVERVIEW OF DATA MINING TOOLS", International Journal of Engineering Applied Sciences and Technology,Vol. 1, Issue 6, ISSN No. 2455-2143, Pages 41-46,2016
2. Hodge, Victoria J., and Jim Austin. "A survey of outlier detection methodologies." Artificial intelligence review, Vol. 22.2, pp.85-126,2004.
3. Chauhan, Prashant, and Madhu Shukla. "A review on outlier detection techniques on data stream by using different approaches of K-Means algorithm." In Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in IEEE, pp. 580-585, 2015.
4. Singh, Karanjit, and Shuchita Upadhyaya. "Outlier detection: applications and techniques." International Journal of Computer Science Issues, Vol. 9.1, pp. 307-323, 2012.
5. Kumar, V. "Parallel and Distributed Computing for Cyber security", Distributed Systems Online, IEEE, pp. 6-12,2005
6. Vaishali, Vaishali. "Fraud Detection in Credit Card by Clustering Approach."International Journal of Computer Applications Vol. 98.3, pp. 29-32, 2014.
7. Kumar, Vijay, Sunil Kumar, and Ajay Kumar Singh. "Outlier Detection: A Clustering-Based Approach." International Journal of Science and Modern Engineering (IJISME), ISSN: 2319-6386, Vol.1,pp. 16-20,2013.
8. https://en.wikipedia.org/wiki/K-means_clustering
9. http://www.math.le.ac.uk/people/ag153/homepage/KmeansKmedoids/Kmeans_Kmedoids.html
10. Hawkins, S., He, H., Williams, G., & Baxter, R. "Outlier detection using replicator neural networks" In International Conference on Data Warehousing and Knowledge Discovery ,Springer Berlin Heidelberg, pp. 170-180, 2002
11. Zhang, Y., Meratnia, N., & Havinga, P. "Outlier detection techniques for wireless sensor networks: A survey", IEEE Communications Surveys & Tutorials, Vol. 12(2), pp. 159-170, 2010.
12. Hodge, V. J., & Austin, J. (2004). "A survey of outlier detection methodologies". Artificial intelligence review, Vol. 22(2), pp. 85-126, 2004.