International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 10 • Issue 2 pp. 211-219 March 2019 - Sept 2019 www.csjournals.com

OUTLIER DETECTION-A STUDY

Ms.S. Maria Sylviaa.S [a], Dr. S.Vijayarani[b]
[a]*Assistant Professor, Department of Computer Science,*
Nirmala college for women, Coimbatore,India.
[b]*Assistant Professor, Department of Computer Science,*
School of Computer Science and Engineering, Bharathiar University, Coimbatore,India

Abstract: outlier detection is used to detect the data which does not fit into the certain pattern. Clustering is an important data mining technique which is used to cluster the data items and also helps to detect the outliers. Outlier is an important problem within various knowledge and applications and it occurs due to human errors, instrumentation faults, fraudulent, etc. Anomalies will be detected through the outlier detection technique. This study paper describes the overview of outlier detection, types of outlier and techniques of outlier which characterizes the different approaches. It also describes various models used for detecting the outliers. Different algorithms used for detecting the outliers are discussed. The challenges give the details about the activities held in outlier and the application of outlier is elucidated about the number of fields where the outlier is detected and how they are detected.
Keywords: Outlier, Types, Models, Techniques, Applications.

1. INTRODUCTION

Outlier detection is a methodof detecting and analyzing the anomalous data which does not fit into a normal behavior. These patterns are referred as noise, outliers, liabilities, novelty, inaccuracy, exceptions, custom, observation defects, contaminants and deviations in distinct application domain [1]. Systematic techniques were used to find the outlier in computer science and statistics. 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 [3].

Outliers are detected in various domains such as master card, medical, intervention [1]. Outliers are fitted into the critical entities like military surveillance where the presence of an unexpected region in a satellite image of an attacker area could indicate an enemy troop movement from a space craft that would be signified a fault [1]. Hawkins formally defined the concept as an outlier is an observation which deviates from the other observations to arouse suspicious event and theseare generated by variousprocedures [2].

Outliers are adapted to the critical entities in military surveillance the presence of an unusual region in a satellite image of enemy area could detect the enemy troop movement or anomalous readings from a space craft which would signify a fault in some component of the craft [1].The outlier detection algorithm is used to detect the anomaly and they are resulted in two ways first, it provides the level of anomaly and they can be used to rank the data points based on their outlierness.

Second, it acknowledges the binary label as an output to find whether they are outlier or not. Outliers are said to be themodel or pattern in the particular data that performs abnormally .The malicious intentions are performed inside the system by the intruders that are detected as outliers and byusing the detection techniques the faults can be detected by monitoring every activities of the data [3].

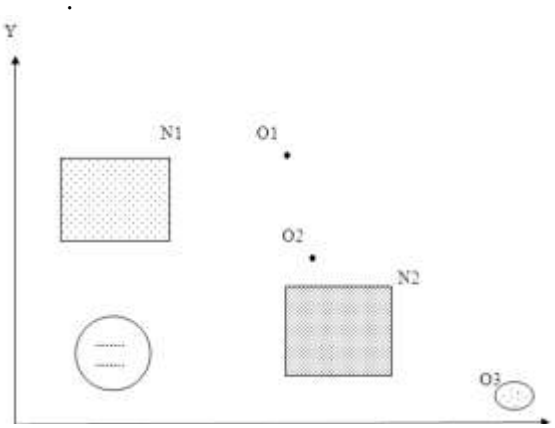

Figure.1. Outlier in 2d data

Figure 1 elucidates outlier in a simple 2d data. The data has two regions in a set which is N1, N2, O1, O2 are the instances which do not meet or lie within the exact region, and O3 is not a normal region. So, O1, O2 and O3 are the outliers.Outlier is

defined as an abnormal event of the data in the particular data set. Almost, every data set has an outlier and detecting outlier the different techniques are used. There are different causes for the outlier.They are

- Malignant activity-fraud activities include credit card, insurance, credit card, cyber intrusion.
- Apparatus error – defects in machine or components.
- Environmental change-occurs due to climatic change, gene mutation, etc.
- Due to wrong data report and accidents humanerrors are occurred.

Outlier detection is related to but distinct from noise removal [9] and noise accommodation [10]. Noise is a data occurrence not in the interest of analyst but data analysis interference. Noise removal is driven by the need to remove the unwanted objects before any data analysis is performed on the data and noise accommodation refers to immunizing statistical model estimation against anomalous observations [4].

Section 2 discusses the types of outlier and how they differ from each other. Section 3 depicts the techniques and the approaches and how to detect the outlier. Section 4 describes about the model and how they are designed. Section 5 gives the popular algorithms used for outlier detection. Section 6 explains about the challenges. Section 7 portrays about the applications. Conclusion is given in Section 8.

2. TYPES OF OUTLIER

Outliers are divided into five types. They are point, trajectory, collective, vector and sequence. They are illustrated in Figure 2.

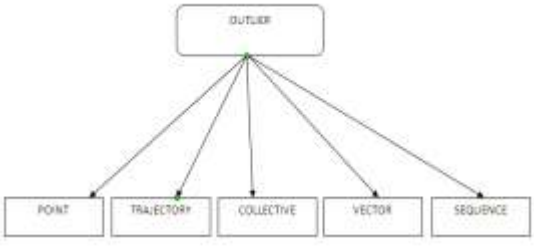

Figure.2.Types of Outlier

Point outliers- In a set of instances the single data instance is defined as point outlier. It is simple and will be detected in most of the existing detection scheme. Every point in the data set is detected as a point outlier rather than with the group of points. The data in a data set is expressed as a vector and every tuple contains number of attributes.

Collective outliers- A collective outlier represents the group ofinterrelateddata instances is anomalous with respect to the whole data set. Data instances which are found in the collective outlier are not considered as outliers by it, but the occurrence of these instancesin a collection is considered as anomalous [1].

Vector outliers- The data in this will look like relational database.The data will be denoted as tuples and each tuple has a set of associated attributes and the data set contains numeric attributes or categorical attributes or both.Data can also beclassified into low dimensional data and high dimensional and even though there is no clear cutoff between these two types of data set [5].

Sequence outliers- The data are denoted as sequence. Example of a sequence database is the computer's system call log where the computer commands are executed in a certain order and they are stored in sequencelike buffer overflow,SMTP [5]. To detect abnormal events the normal eventsare maintained and the abnormalities are labeled as sequence.

Trajectory outliers-Trajectory data can be collected for moving objects such as satellites and vehicle. Examples of trajectory are vehicle positioning data, hurricane tracking data, and animal movement data [6]. Trajectory data is expressed as a set of keys along with start and end points, min value, max value, average, min velocity and max velocity. Based on this representation adequate sum of distance function can be defined to compute the difference of trajectory based on the key features [7]. The recent work proposed a partition and designed a framework for detecting trajectory outliers [6]. The outlier can be found by segmenting the data into line segment instead of whole data.

Graph outliers- Graph outliers define the entities of graph which is acted abnormally. The entity of graph which becomes the outlier is edges and nodes. Auto part detects outlier edges in a general

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 10 • Issue 2 pp. 211-219 March 2019 - Sept 2019 www.csjournals.com

graph [8]. The outliers are denoted as vector data set.

3. OUTLIER DETECTION TECHNIQUES

A technique of outlier detection defines about the type which is used by training data set for the detection.

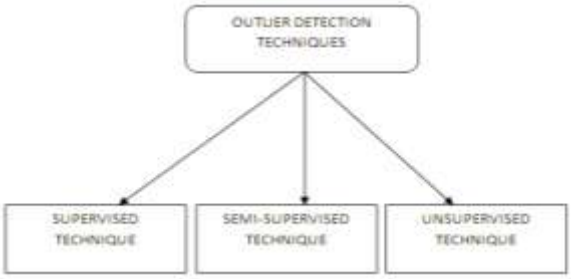

Figure.3.Outlier Technique

3.1 Supervised technique

Techniques trained in supervised mode assume that the availability of a training data set which has labeled instances for normal as well as outlier class [16].The class can be determined by comparing the data instances. There are different problems arise in supervised technique they are

In the training data set the outlier instances are very far from the normal instances due to this problem, i.e. the classes are imbalanced.

Obtaining exact and representative labels for the anomaly class is usually challenging and number of techniques has been proposed that injects the artificial anomalies in a normal data set to obtain a labeled training data set [16].

3.2 Semi-Supervised technique

Techniques that operate in a semi- supervised mode assume that the training data has labeled instances and these labeled instances are not considered for outlier detection. [16].Semi-supervised mode is more effective than supervised mode because the label is not required for the outlier class. A model is built for the class which is comparable to normal characteristic of the data and the outlier is found using test data.This type of techniques is not commonly used because it is very difficult to access the training data set and the data will behave in anomalous way. Assumptions in semi-supervised techniques are

Smoothness assumption: The labels are shared by the points which are very adjacent to each other.

Cluster assumption: Distinct clusters are formed and the labels share the points among the equivalent clusters.

Manifold assumption: In the input space the data lies on the lower dimension. It is very difficult to model the high dimensional data. To avoid the curse of dimensionality the labeled and unlabeled data are used for the literature of manifold.

3.3 Unsupervised technique

An unsupervised technique does not require training data and they are most widely used. The normal instances are far than the outliers in the test data [16].If the assumed result does not appear then the high false alarm is set. Many semi-supervised techniques can be adapted to operate in unsupervised mode by using a sample of the unlabeled data set as training data [16]. According to the assumption the test data contains minor outlier.Approaches to unsupervised learning include

- clustering
- hidden Markov models,
- Feature extraction techniques

4. MODELS IN OUTLIER DETECTION

There are various models used in outliers. The interpretability of an outlier detection model is extremely important from the perspective of the analyst [2]. It is often desired to regulate a specific point as an anomaly in terms of its performance with other data. It provides the hints to the analyst about the diagnosis required in an application specific scenario and this is also referred as the intentional knowledge about the outliers [11]. Figure 4 displays the different outlier detection models.

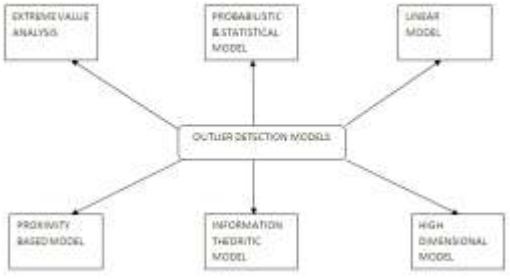

Figure.4.Outlier detection models

4.1 Extreme value analysis

Extreme value analysis is one of the basic forms of outlier detection for one dimensional data. This type of outlier is simulated as when the value is too high or too low. The key is to determine the statistical tails of the underlying distribution and the normal distribution and it will be the easiest thing to analyze because most statistical test can be interpreted directly in terms of probabilities of significance [2]. The z value test brings the outlier point of data to be valuable. Hawkins defines the objects by their generative probabilities rather than the extremity values [2]. For example, in a data set {161, 1, 114, 50, 98, 8, 4} with one-dimensional data where 161 and 114 are determined as extreme values so, it is considered as the outlier.

Confusions between extreme value analysis and outlier analysis are common especially in the context of multivariate data and extreme value analysis is naturally designed for univariate data [2].The modeled deviation of extreme value analysis is done as the last step and they are denoted as univariate score and the high value is taken as outlier. Using the distance or depth method the multivariate data can be specified.

4.2 Probabilistic and Statistical Models

Probabilistic and statistical models are considered to be the oldest methods and these were intended without much focus on realistic issue like efficiency and representation of data. For example, a Gaussian mixture model is a generative model which characterizes the data in the form of a productive process which contains a mixture of Gaussian clusters [2].EM algorithm is applied to the data set to find the parameters of Gaussian distribution and the result is that the chance of data points to distinctive clusters. Here, the low qualified data points are denoted as outlier. A major advantage of probabilistic models is that they can be virtually applied to any type of data as long as an appropriate for each mixture component [2].Every result regarding this model will be displayed as statistical view. For a mixture of divergent types of attributes a product of the attribute will be specific to the generative components [2]. A downside of probabilistic models is that they will try to fit the data into particular distribution which is not appropriate for that data.

4.3 Linear model

With the help of linear correlation the data is modeled into low dimensional fixed sub space. Extreme values analysis can be applied on these deviations in order to determine the outliers [2]. The concept of dimensionality reduction and principal component analysis which is denoted as PCA is quite similar except that it uses a non-parametric approach in order to model the data correlations [12]. PCA can be derived through multivariate regression analysis by determining the plane which optimizes the least square error of representation in terms of the normal distance to the plane [2]. The outlier can be defined by the correct size of the residual.Data points which have large residuals are more likely to be outliers because their behavior does not conform to the natural subspace patterns in the data[2]. To reduce the noise of the data in the data set PCA is used. The embedded subspace is said to be the linear combination of features with true or false coefficient in which true represents positive and false as negative.

4.3.1 Spectral Models

The matrix decomposition used in the background of graphs is spectral and they are defined as the model which is used to define about the matrix creation. PCA is one of the matrix decomposition method used in the framework of graphs and networks. Spectral methods are used commonly for clustering graph data sets and are also used to identify anomalous changes in temporal sequences of graphs [13].

4.4 Proximity based model

In this proximity model the outliers are denoted as a point and they are differentiated from the remaining data.The modeling of the data is done by three methods namely

Cluster analysis

Density based

Nearest neighbor.

In clustering analysis and other density based methods the denser region in the data are found directly and outliers are defined as those points which do not lie on these denser part [2].The contrast between the density and clustering analysis is that the clustering will segment the points and density will segment the space. In nearest neighbor method the distance of each data point to its kth nearest neighbor is determined [2].

When applying the clustering method first the algorithm is used to find the dense region. Second, it try to fit the data point into the number of clusters to produce the value for the outlier. For example, in

the case of k-means clustering algorithm the distance of the data point to the nearest centroid may be used as a measure of its anomalous behavior [2].Density based methods provide a high level of interpretability when the sparse regions in the data can be presented in terms of combinations of the original attributes [2].

4.5 Information theoretic model

In order to perform the analysis, all the above mentioned models have used summarization, modeling, detecting the outlier, visualizing in the form of graph and probabilitytechniques. This information theoreticmodel provides a small summary of the data in which the deviated data is flagged and these are denoted as outliers [2].The code length is increased in the data set. For example,

JKJKJKJKJK

JKJKSKJKJK

Here, both the strings are equal and in the second string the S is the character used single time. In the first string JK is used 5 times and the new character isintroduced in the second string and it will increase the lower limit of the length. The character which is different in the string is treated as an outlier. Information theoretic models are closely related to conventional model because both used a concise representation of the data set which is a baseline for comparison [2]. A probabilistic model describes a data set in terms of generative model parameters, such as a mixture of Gaussian distributions or a mixture of exponential power distributions [14].

4.6 High dimensional model

In high dimensionality the data becomes sparse and all pairs of data points become almost equidistant from one another so this case is challenging [15]. Dimensionally all regions become sparse in density model. The sparsity behavior in high dimensionality makes all points look very similar to one another and the observation is true and these outliers may be discovered by examining the distribution of the data in a lower dimensional local subspace [2].Low dimensional subspace contains outlier and the character is latent to analysis. This is the combinational process of clustering and regression. To show the outlier pattern analysis and subspace analysis high dimensional model is determined. This can be a huge challenge to compute the high dimensional data because the simultaneous discovery of relevant data localities and subspaces is very difficult [2].High-dimensional methods provide an interesting direction for intentional understanding of outlier analysis when the subspaces are described in terms of the original attributes [2].

5.ALGORITHMS FOR OUTLIER

5.1 PAM ALGORITHM (Partition around medoids)

PAM is a powerful method here the medoids are used to perform partitioning and detecting outliers whereas in k-means centroids are used for partitioning. [17].Medoid isreferred as the representative object of a data set. There are two phases in PAM01

A).**Build phase:** It selects the k objects which are centrally located and they can be referred as k medoids [17].

B)**Swap phase**: The total cost is calculated for every pair which is a combination of selected and non selected objects.

The dataset D is taken as input and consider the medoids which randomly selects the k objects and calculates the total cost separately for each pair of selected and non selected object. If the total cost of selected pair of objects is less than 0 then the pair is replaced as non selected pair. Then the new medoid is taken for the non selected object. The problem with the PAM algorithm is that when there is no difference in the total cost T for each pair then the algorithm stops its execution.

PAM Pseudo code

1. Input the dataset D
2. Randomly select k objects from the dataset D
3. Calculate the Total cost T for each pair of selected S_i and non selected object S_h
4. For each pair of T if S_i <0, then it is replaced with S_h
5. Then find similar medoid for each non-selected object
6. Repeat step 2, 3 and 4 until finding medoids

5.2 CLARA (Clustering large applications)

To overcome the problem of PAM, CLARA is used. It works efficiently in larger databases. It takes only the sample datum from the dataset instead of full dataset [17].It selects the data from the set and finds medoids for finding the outlier. The dataset D is

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 10 • Issue 2 pp. 211-219 March 2019 - Sept 2019 www.csjournals.com

taken as input and the sample data is taken as S where the PAM concept is applied to CLARA to find the medoidM. After finding the medoid the entire dataset D is classified based on cost then the average is calculated to find the dissimilarities between the clusters.

CLARAPseudo code

1. Input the dataset D
2. Repeat n times
3. Draw sample S randomly from D
4. Call PAM from S to get medoids M
5. Classify the dataset D to $cost_1$.....$cost_k$
6. Calculate the average dissimilarity from the obtained clusters

5.3 LOCAL OUTLIER FACTOR

This algorithm isproposed by Markus M. Breunig, Hans-Peter Kriegel [18], Raymond T. Ng and Jorg Sander in 2000 for finding anomalous data points by measuring the local deviation of a given data point with respect to its neighbors [18]. Higher amounts of data are collected and they are stored in databases by increasing the need of the data for efficient and effective analysis methodsto make use of the information contained implicitly in the data[18].LOF is found to the every data in the dataset to indicate the degree of outlierness. The outlier factor is local in the sense that only a restricted neighborhood of each object is taken into an account [18].The reachabilitydistance should be found from the centroid to define the outlierness.

$$\text{reach-dist}_k(P, O) = \max\{\text{k-distance}(O), d(P, O)\}$$

Here,dist is the distance and P,O are the objects where thereachability distance is calculated from the object P to O which belongs to the k nearest neighbor of O.From this the local density is found for the single object P.

$$lrd_{MinPts}(p) = 1 / \left(\frac{\sum_{o \in N_{MinPts}(p)} reach\text{-}dist_{MinPts}(p, o)}{|N_{MinPts}(p)|} \right)$$

The local reachability density(lrd) of an object pis the inverse of the average reachability distance based on the MinPtsnearest neighbors of p.The LOF is found for the point P by the formulae.

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|}$$

Here the LOF is found for the minimum points (MinPts)for the object P. It is the average of the ratio of the local reachability density of p and those of p's MinPts-nearest neighbors [18].

6. CHALLENGES

The challenge in the outlier detection is that it involves exploring the unseen space and an outlier can be defined as a pattern which does not fitinto an expected normal behavior [1].

Several factors have been described for the outlier detection and they are very simple.

- Every normal region is defined and contains abnormal behavior which is very difficultto interpret.
- Normal behaviors have been derived and actual approach is not sufficient for the future.
- The boundary between normal and outlying behavior is often fuzzy and thus an outlying observation which lies close to the boundary can be actually normal [1].
- Different domains have different approaches and all the domains have set of requirements and rules and have a formulation for the outlier detection.
- While developing an outlier detection technique the labeled data for validating and for training has major issues. It is very difficult to organize the ordered list of objects behavior.
- In several cases the outliers are the result of malicious actions and malicious adversaries that adapt themselves to make the outlying observations appear like normal thereby making the task more difficult [1].
- Handling noise in the dataset is the difficult process. The noise is detected in every data and the noise is similar to the outlier and it is more difficult to eliminate and they distort the object behavior and reduce the effectiveness.
- Interpretability of the outlier score are found by the better performance or better normalization. Specific formulation is focused to detect the outlier in the data set and they are very difficult to work on with**.** The formulation is induced by two factors they are, nature of data and nature of outliers [1]**.**A general

challenge in outlier detection is not limited to improve the ensemble methods for outlier detection. It is very important to find the internal methods of detecting the outlier and the criteria should be evaluated by themselves. Aggregating the rank of outlier of different subset of dimensions is defined as feature bagging.

Requirements of the outlier detection

1. Nature of data, nature of outliers, other constraints and assumptions collectively constitute the problem formulation [1].
2. Application domain in which the technique is applied and some of the techniques are developed in a more generic fashion but are still feasible in one or more domains while others directly target a particular application domain[1].
3. The concept and ideas used from one or more knowledge disciplines [1].

7. APPLICATIONS

There are numerous applications that are used to detect outlier. This defines about the applications and how they can be detected.

7.1 Intrusion Detection System

Intrusion detection refers to detect the malicious activity in a computer related system [1]. From the computer perspective intrusion detection is an important task**.** Outlier detection techniques have been widely applied for intrusion detection and the key challenge for detecting the outlier in this domain with huge volume of data is a difficult process [1].The detection scheme should be computed efficiently to maintain large inputs. Thelabeled data corresponding to normal behavior is readily available where labels for intrusions are harder to obtain for training. Intrusions are classified into two ways they are, host based intrusion and network based intrusion

7.1.1 Host based IDS

These systems are also named as system call intrusion detection systems which deals with system calls. These calls could be generated by programs or by the data is sequential in nature and the alphabet consists of individual system calls like the alphabet is usually small [1]. The programs are executed by the system calls with its own sequence.

7.1.2 Network based IDS

Detection of intrusion can be done in network area. A typical setting is a large network of computers which is connected to the rest of the world via the Internet. Here, the system works with different level of data**.** The data has a temporal aspect associated with it but most of the applications typically do not handle the sequential aspect explicitly.

7.2 Fraud detection

It is the criminal action occurring in the organizations like stock market, insurance, banks, etc.The malignant person acts like a customer and retrieve the organization's information in unauthorized ways. The organizations are interested in immediate detection of such frauds to prevent economic losses [1]. Fawcett and Provost have introduced the term activity monitoring which is a general approach for detecting frauds in these domains [1]. The outlier technique is used to maintain and use the profile of every customer and views to detect the anomaly.

7.2.1 Credit Card Fraud Detection

This is a very important domain where credit card companies are interested in either detecting fraudulent credit card applications or fraudulent credit card usage [1].The credit card frauds can also be detected in online usage. The problem in this fraudulent area can be addressed by two different ways of usingthe outlier detection techniques. The first problem is that every owner of the credit card has their own profile and every transaction is correlated with the profile and if the outlier is detected then they are flagged. This approach is typically expensive since it requires querying a central data repository every time a user makes a transaction [1]. Second problem is found by the operation that detects the anomaly in specific location transaction. Thus in this approach the algorithm is executed locally and hence it is less expensive [1].

7.2.2 Mobile Phone Fraud Detection

Mobile or cellular fraud detection is an activity monitoring problem and the task is to scan a large set of accounts examining the calling behavior of each user and to issue an alarm when an account appears to be defrauded [1]. For example

the call hours or call days of a user or area of the user can be detected depending on the granularity.

7.3 Medical and Public Health Data

The patient's record are collected to detect the outlier in medical and public health domain.The data can have outliers due to several reasons such as abnormal patient condition or instrumentation errors or recording errors and the outlier detection is a very critical problem in this domain and requires high degree of accuracy [1]. The patient's record contains different types of attribute such as name, age, weight, etc.The data might also have temporal as well as spatial aspect to it and most of the current outlier detection schemes aim at detecting outlying records.

7.4 Industrial Damage Detection

Industrial damage occurs due to the regular usage. The damages should be detected and prevented to avoid losses using outlier detection techniques. The data referred as sensor data because it is recorded using divergent sensors and collected for analysis [1].

7.5 Image Processing

Outlier detection techniques dealing with images are either interested in any changes in an image over time or in regions which appear abnormal on the static image [1].The images of satellite, digit recognition are the examples. Different techniques are used in image processing to avoid anomaly. Every data point has attributed like color, texture. Outliers are detected by the incoming of objects, error in instrumentation.One of the key challenges in this domain is the large size of the input and hence computational efficiency of the outlier detection technique is an important issue [1].

8. CONCLUSION

Outlier detection is one of the important tasks and it is performed in numerous applications. In this study paper many numbers of techniques and approaches have been discussed.Analyzing the anomaly in unviewed space is the key challenge of outlier detection. Outlier is an important problem within various knowledge and applications. There are numerous techniques available to detect outlier and every techniques have their own limitations. This paper provides the overall study about the outlier detection with the data mining perspective. The effectiveness of detecting outlier is dependent on the type of data and the method chosen. New techniques can be proposed to detect outlier and it can be applied in different domains in future.

REFERENCE

[1] Outlier Detection: A Survey-varun chandola, arindam banerjee and VipinKumar, TR 07- 017 August 15, 2007

[2] Outlier analysis-Charu.c.aggarwal, BM T.J.Watson Research Center, Yorktown Heights, NY, US, Kluwer Academic Publishers

[3] A Survey of Outlier Detection Methodologies-Victoria J. Hodge and Jim Austin, Springer 2004

[4] Outlier Detection: Applications And Techniques-Karanjit Singh and Dr. Shuchita Upadhyaya, IJCSI, International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012 ISSN (Online): 1694-0814

[5] Advancements of Outlier Detection: A Survey-Ji Zhang, ICST Transactions on Scalable Information Systems January-March 2013, Volume 13, Issue 01-03

[6] J. Lee, J. Han and X. Li.-Trajectory Outlier Detection: A Partition-and-Detect Framework, ICST Transactions on Scalable Information Systems, January-March 2013, Volume 13.

[7] E.M.Knorr, R.T. Ng andV. Tucakov. Distance-Based Outliers: Algorithms and Applications, Springer-Verlag New York, Inc. Secaucus, NJ, USA, Volume 8, Issue 3-4, February2000, Pages237-253

[8] D. Chakrabarti-Auto part: Parameter-free graph partitioning and outlier detection, Springer, Lecture Notes in Computer Science Volume 3202, 2004, pp 112-124,

[9] Teng, H. Chen, K. and Lu, S. 1990. Adaptive real-time outlier detection using inductively generated sequential patterns. InProceedingsofIEEE Computer Society Symposium on Research in Security and Privacy. IEEE Computer Society

[10] Rousseau, P. J. and Leroy, A. M. 1987. Robust regression and outlier detection, Wiley Interscience, New York (Series in Applied Probability and Statistics), 329 pages. ISBN 0-471-85233-3.

[11] E. Knorr, and R. Ng. Finding Intentional Knowledge of Distance- Based Outliers, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA ©1999, Pages 211-222, ISBN:1-55860-615-7

[12] I. Jolliff e-Principal Component Analysis, Springer Series in Statistics, 2nd ed. 2002, Page 488

[13] T. Ide and H. Kashima-Eigen space based Anomaly Detection in Computer Systems, ACM KDD Conference 2004.

[14] C. Bohm, K. Haegler, N. Muller, and C. Plant-Coco: Coding Cost for Parameter Free Outlier Detection, ACM KDD Conference 2009.

[15] L. Kaufman and P. Rousseeauw- Finding Groups in Data:-An Introduction to Cluster Analysis, Wiley Interscience 1990.

[16] Anomaly Detection: A Survey-varun chandola, arindam banerjee, VipinKumar, ACM Computing Surveys (CSUR), Article No. 15 Volume 41 Issue 3, July 2009

[17]An Efficient Clustering Algorithm for Outlier Detection S.Vijayarani S.Nithya International Journal of Computer Applications (0975 – 8887) Volume 32– No.7, October 2011.

[18] Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. (2000). "LOF: Identifying Density-based Local Outliers". Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data.