Handling Outliers Efficiently Using Partition Based Clustering Techniques

Tarrnum Khan[1] Mr. Ranjan Thakur[2]
[1]P.G. Research Scholar [2]Assistant Professor
[1,2]Department of Computer Science Engineering
[1,2]JIT Borawan Khargone, India

Abstract— Outliers are data which can be considered anomalous due to several causes (e.g. erroneous measurements or anomalous process conditions). Outlier detection techniques are used, for instance, to minimize the influence of outliers in the final model to develop, or as a preliminary pre-processing stage before the information conveyed by a signal is elaborated. The traditional outlier detection methods can be classified into four main approaches: distance-based, density-based, clustering- based and distribution-based. In the proposed approach is based on two partition based clustering methods K-Mean and PAM. K Mean method is based on mean values of the object belongs in the cluster. K- mean find the distance of each objects from the mean objects and the object which has minimum distance from the mean object are keep in the cluster otherwise the object are swapped into the cluster whose mean has minimum distance. Some time when outliers are present in the data set we consider these outliers in any one of the cluster. These outliers are affecting the mean value of the clusters. Because of these outliers which are basically not belongs to any of the clusters we have to consider in then ion K-mean. But in case of PAM we choose the objects as the medoid as a cluster center, so when we calculate the distance of other objects with this center objects ,the outliers are easily identified.

Keywords: K-Mean, PAM, Clustering Techniques

I. INTRODUCTION

Outliers are generally defined as samples that are exceptionally far from the mainstream of data. There is no rigid mathematical definition of what constitutes an outlier; determining whether or not an observation is an outlier is ultimately a subjective exercise. An outlier may also be explained as a piece of data or observation that deviates drastically from the given norm or average of the data set. An outlier may be caused simply by chance, but it may also indicate measurement error or that the given data set has a heavy-tailed distribution. Therefore, Outlier Detection may be defined as the process of detecting and subsequently excluding outliers from a given set of data. There are no standardized Outlier identification methods as these are largely dependent upon the data set. Outlier Detection as a branch of data mining has many applications in data stream analysis[10,11].

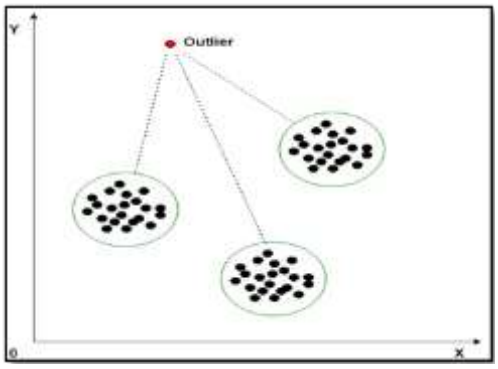

Fig. 1: Outlier in red colour

Outlier detection is an important branch in data mining, which is the discovery of data that deviates a lot from other data patterns. Data mining is primarily used today by companies with a strong consumer focus retail, financial, communication, and marketing organizations. It is used to determine relationships among the internal factors such as price, product positioning, or staff skills, and external factors, such as economic indicators, competition, and customer demographics. Also, it enables them to determine the impact on sales, customer, satisfaction, and corporate profits. Data mining related methods are often non-parametric, thus, it does not assume an underlying generating model for the data. These methods are designed to manage the large databases from high-dimensional spaces. The identification of outliers can lead to the discovery of unexpected knowledge in areas such as calling card fraud detection, credit card fraud detection, discovering criminal behaviors, computer intrusion detection, etc. Applications such as outlier detection network intrusion detection, customized marketing, weather prediction, pharmaceutical research and exploration in science databases require the detection of outliers. Outlier detection is an important branch in data pre-processing and data mining, as this stage is required in elaboration and mining of data coming from many application fields such as industrial processes, transportation, ecology, public safety, climatology [12,13]

II. TYPES OF OUTLIERS

In general, outliers can be classified into three categories, namely global outliers, contextual (or conditional) outliers, and collective outliers. An important aspect of an outlier detection technique is the nature of the desired outlier. Outliers can be classified into following three categories.[14,15]

A. Global Outliers

In a given data set, a data object is a global outlier, if it deviates significantly from the rest of the data set. Global outliers are sometimes called point anomalies, and are the simplest type of outliers. Most outlier detection methods are

aimed at finding global outliers, To detect global outliers, a critical issue is to find an appropriate measurement of deviation with respect to the application in question. Various measurements are proposed, and, based on these, outlier detection methods are partitioned into different categories. Global outlier detection is important in many applications.

B. Contextual Outliers

In a given data set, a data object is a contextual outlier if it deviates significantly with respect to a specific context of the object. Contextual outliers are also known as conditional outliers because they are conditional on the selected context. Therefore, in contextual outlier detection, the context has to be specified as part of the problem definition. Generally, in contextual outlier detection, the attributes of the data objects in question are divided into two groups.

C. Collective Outliers

Given a data set, a subset of data objects forms a collective outlier if the objects as a whole deviate significantly from the entire data set. Importantly, the individual data objects may not be outliers. Global or contextual outlier detection, in collective outlier detection we have to consider not only the behavior of individual objects, but also the groups of objects. Therefore, to detect collective outliers, we need background knowledge of the relationship among data objects such as distance or similarity measurements between objects.

III. CAUSES OF OUTLIERS

Anscombe & Guttman (1960), had attempted to categorize the different ways in which outliers may arise. It was relevant to consider them in rather more detail. In taking observations, different sources of variability can be encountered. We can distinguish three of these. [16,17]

A. Natural variability

This is the expression of the way in which observations intrinsically vary over the population; such variation is a natural feature of the population and uncontrollable. Thus, for example, the measurements of heights of men will reflect the amount of variability indigenous to that population.

B. Measurement errors

Often we must take measurements on members of a population under study. Inadequacies in the measuring instrument superimpose the further degree of variability on the inherent factor. The rounding of obtaining values or mistakes in recording compound the measurement error: they are part of it. Some control of this type of variability is possible.

C. Execution error

A further source of variability arises in the imperfect collection of our data. We may inadvertently choose a biased sample or include individuals who are not truly representative of the population we aimed to sample. Again, sensible precautions may reduce such variability

IV. APPLICATIONS OF OUTLIER DETECTION

Outlier's detection can be applied on lot of data sets for various purposes. Some of which are discussed below.[18,19]

- Fraud detection – Detecting fraudulent applications for credit cards, state benefits or detecting fraudulent usage of credit cards or mobile phones. Fraud refers to criminal activities occurring in commercial organizations such as banks, credit card companies, insurance agencies, cell phone companies, stock market, etc. Malicious users could be actual customers of the organization or resorting to identity theft (posing as customers). The detection activity aims at detection of unauthorized consumption of resources provided by the organization to prevent economic losses.
- Fraudulent Usage of Credit Card: Associated with credit card thefts. The data records are defined over several dimensions such as the user ID, spent amount, time between consecutive card usage, etc. The frauds are typically reflected in transactional records (point outliers) and correspond to high payments; high rate of purchase, purchase of items never purchased by 27 the user before, etc. Availability of labeled records is no problem since credit companies have complete data available. Moreover, the data falls into distinct profiles based on the credit card user. Hence profiling and clustering based techniques are typically used in this domain.
- Intrusion detection- Detecting unauthorized access in computer networks. Intrusion detection refers to detection of malicious activity (break-ins, penetrations, and other forms of computer abuse) in a computer related system interesting from a computer security perspective. Being different from normal system behavior, intrusion detection is a perfect candidate for applying outlier detection techniques.
- Labeled data not usually available for Intrusions: This gives preference to semi supervised and unsupervised outlier detection techniques. Intrusion detection systems have been classified into host based and network based intrusion detection systems
- 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.
- Structural defect detection - monitoring manufacturing lines to detect faulty production runs for example cracked beams.
- Satellite image analysis - identifying novel features or misclassified features.
- Detecting novelties in images - for robot neo taxis or surveillance systems.
- Motion segmentation - detecting image features moving independently of the background.
- Time-series monitoring -monitoring safety critical applications such as drilling or high-speed milling.
- Medical condition monitoring - such as heart-rate monitors.

V. LITERATURE SURVEY

1) In 2011 Juliano Gaspar proposed "A Systematic Review of Outliers Detection Techniques In Medical Data" Background: Patient medical records contain many entries relating to patient conditions, treatments and lab results. Outlier detection techniques can be used to detect abnormal patterns in health records (for instance, problems in data quality) and this contributing to better data and better knowledge in the process of decision making. The literature was systematically reviewed to identify articles mentioning outlier detection techniques or anomalies in medical data. Four distinct bibliographic databases were searched: Medline, ISI, IEEE and EBSCO. 4071 distinct papers selected, 80 were included after applying inclusion and exclusion criteria. According to the medical specialty 32% of the techniques are intended for oncology and 37% [1]
2) In 2012 Karanjit Singh and Dr. Shuchita Upadhyaya proposed "Outlier Detection: Applications And Techniques" Outliers once upon a time regarded as noisy data in statistics, has turned out to be an important problem which is being researched in diverse fields of research and application domains. Many outlier detection techniques have been developed specific to certain application domains, while some techniques are more generic. Some application domains are being researched in strict confidentiality such as research on crime and terrorist activities.. Outlier detection aims to find patterns in data that do not conform to expected behavior. It has extensive use in awide 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 [2].
3) In 2013 Jyoti Ranjan proposed "Study of Distance-Based Outlier Detection Methods" An Outlier is an observation which is different from the others in a sample. Usually ananomaly occurs in every data due to measurement error. Anomaly detection is identifying anomalous data for given dataset that does not show normal behavior. The outlier detection can be grouped into 5 main categories: statistical-based approaches, depth based approaches, clustering approaches, distance-based approaches and density-based approaches. Distance - based methods i.e. Index-based algorithm, Nested-loop algorithm and LDOF are discussed. To reduce the false positive error in LDOF, They proposed MLDOF algorithm. They tested LDOF and MLDOF by implementing on several large and high-dimensional real datasets obtained from UCI machine repository. The experiments show that the MLDOF improves accuracy of anomaly detection with respect to LDOF and reduces the false positive error [3].
4) In 2014 Manish Gupta, Jing Gao proposed "Outlier Detection for Temporal Data: A Survey" In the statistics community, outlier detection for time series data has been studied for decades. Recently, with advances in hardware and software technology, there has been a large body of work on temporal outlier detection from a computational perspective within the computer science community. They presented an organized overview of the various techniques proposed for outlier detection on temporal data. Modeling temporal data is a challenging task due to the dynamic nature and complex evolutionary patterns in the data. In the past, there are a wide variety of models developed to capture different facets in temporal data outlier detection. This survey organized the discussion along different data types, presented various outlier definitions, and briefly introduced the corresponding techniques. Finally, they discussed various applications for which these techniques have been successfully used [4].
5) In 2015Usman Habib, Gerhard Zucker proposed "Outliers Detection Method Using Clustering in Buildings Data". They discuss the steps involved for detecting outliers in the data obtained from absorption chiller using their On/Off state information. It also proposes a method for automatic detection of On/Off and/or Missing Data status of the chiller. The technique uses two layer K-Means clustering for detecting On/Off as well as Missing Data state of the chiller. After automatic detection of the chiller On/Off cycle, a method for outlier detection is proposed using Z-Score normalization based on the On/Off cycle state of chillers and clustering outliers by Expectation Maximization clustering algorithm.. The two layered K-Means algorithm gives 3 states of a chiller On/Off and Missing Data state. The Missing Data state is representing the duration when there is no data recorded for any sensor of the chiller [5]
6) In 2016Kamaljeet Kaur proposed "Comparative Study of Outlier Detection Algorithms" As the dimension of the data is increasing day by day, outlier detection is emerging as one of the active area of research. They covers a study of various outlier detection algorithms like Statistical based outlier detection, Depth based outlier detection, Clustering based technique, Density based outlier detection etc. Comparison study of these outlier detection methods is done to find out which of the outlier detection algorithms are more applicable on high dimensional data. The speed of processing the data is to be increased that helps in the reduction of processing cost of data. There is no single universally applicable outlier detection approach of the current techniques. They present the study of different existing outlier detection techniques and the way in which they are categorized. It is concluded that performance of clustering algorithms is comparatively better than other outlier detection algorithms on huge data sets [6].
7) In 2017 Rasim M. Alguliyev, Ramiz M. Aliguliyev proposed "An Anomaly Detection Based on Optimization ". At present anomaly detection is one of the important problems in many fields.. They improved optimization approach for a previously known number of clusters, where a weight is assigned to each data point, is proposed. Their aim is to show that weighting of each data point improves the clustering solution. The quality of the clustering result was estimated using clustering evaluation metrics. They showed that the proposed method works better than k-means on the Australia (credit card applications) dataset according to the Purity,

Mirkin and F-measure metrics, and on the heart diseases dataset according to F-measure and variation of information metric. They show that weighting improves the clustering solution. The comparison was made using three data sets containing anomalous values. The quality of the clustering result was estimated using six metrics [7].

8) In 2018 Victoria J. Hodge and Jim Austin proposed "An Evaluation of Classification and Outlier Detection Algorithms". They evaluated algorithms for classification and outlier detection accuracies in temporal data. They focus on algorithms that train and classify rapidly and can be used for systems that need to incorporate new data regularly. They compare the accuracy of six fast algorithms using a range of well-known time-series datasets. The analyses demonstrate that the choice of algorithm is task and data specific that derive heuristics for choosing. Gradient Boosting Machines are generally best for classification but there is no single winner for outlier detection though Gradient Boosting Machines (again) and Random Forest are better. They recommend running evaluations of a number of algorithms using our heuristics. They evaluated algorithms for both classification and outlier detection for an on-line system that assimilates new data regularly. They aimed to derive heuristics for the best algorithms [8].
9) In 2019Yue Zhao proposed PyOD: A Python Toolbox for Scalable Outlier Detection PyOD: A Python Toolbox for Scalable Outlier Detection PyOD is an open-source Python toolbox for performing scalable outlier detection on multivariate data. Uniquely, it provides access to a wide range of outlier detection algorithms, including established outlier ensembles and more recent neural network-based approaches, under a single, well-documented API designed for use by both practitioners and researcher. PyOD is compatible with both Python 2 and 3 and can be installed through Python Package Index (PyPI). They presented PyOD, a comprehensive toolbox built in Python for scalable outlier detection. It includes more than 20 classical and emerging detection algorithms and is being used in both academic and commercial projects. They planned to enhance the toolbox by implementing models that work well with time series and geospatial data, improving computational efficiency through distributed computing and addressing engineering challenges such as handling sparse matrices or memory limitations [9].

VI. PROBLEM STATEMENT

As lot of outlier detection algorithms exists for detecting outliers and the usage of all these vary according to the type.

1) Efficiency:-It is the evaluation of the average execution time required for an algorithm to complete work on a given data set. Efficiency of an algorithm is measured by its order. It is helpful for quantifying implementation difficulties of certain problems.
2) Computational Cost: - It is directly proportional to the computational complexity of the algorithm. It is the valuation of the number of steps required by the algorithm related for input of an instance or a given size in the worst case. Function of size is measured by the number of steps.
3) Scalability: - It is defined as the capability of the product or a computer application to continue to function well even when it is changed in size or volume, as per the user requirements. It is basically a rescaling like expandability of an application program which can be used on larger operating systems for handling large number of users and also for better performance
4) Applicability:-
   As each algorithm has its boundaries and limits set for being applicable on any given set of data. Depending upon the data set i.e. whether it's a statistical data or large dataset, various algorithms are applied on the datasets to detect outliers. All the above stated outlier detection algorithms are compared in table-1 with respect to certain parameters like efficiency, computational cost, scalability, applicability etc.

VII. K MEAN CLUSTERING METHOD

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labeled, outcomes.

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

K-Means algorithm works is as follows

Let $X = \{x_1, x_2, \ldots, x_n\}$ represent set of objects and $V = \{v_1, v_2, \ldots, v_c\}$ represent set of centers.

1) Select number of clusters K.
2) Randomly select objects for without replacement and find mean for each cluster and decide as center.
3) Calculate distance between objects and center.
4) Allocate each object to the closest centroid.
5) Update center after allocating objects.
6) Keep iterating until there is no change to the center and no object move from clusters.
7) End

VIII. PAM CLUSTERING METHOD

PAM algorithm slightly modified version of K-Means algorithm. In K-Means algorithm means is choose as the centroids of the clusters but in case of PAM centers are selected in such a way that all other objects are evenly distributed around centers. Both the *k*-means and PAM algorithms are partitioned methods. K-means attempts to minimize the total squared error, while PAM minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster. PAM is also a partitioning technique of clustering that clusters the data set of n objects into k clusters with k known a priori. A medoid of a finite dataset is a data point from this set, whose average dissimilarity to all the data points is minimal i.e. it is the most centrally located point in the set. The most

common realization of k-medoid clustering is the Partitioning Around Medoids (PAM) algorithm.

A. Algorithm of PAM clustering

PAM algorithm has following steps

1) Initially select K random points as the centers from the given n objects of the data set.
2) Associate each objects to the closest center by using distance metrics.
3) Calculate total swapping cost for pair of non-selected and selected objects.
4) Calculate total cost by adding individual cost of each clusters. If the total costs of constructed clusters is minimum as compared to the all previous cost then end clustering process.
5) Repeat the steps 2-3 until there is no change of the center.

IX. Result and Analysis

We evaluate the performance of K Mean and PAM clustering methods and compare with number of outlier generated by both approaches. The experiments were performed on Intel Core i3processor1GB main memory and RAM: 4GB Inbuilt HDD: 400GB OS: Windows7. The algorithms are implemented in using Dot Net Framework language version 4.0.1. Synthetic datasets are used to evaluate the performance of the algorithms.

We have taken weight and height of more than 100 students. We used K Means and PAM method to construct clusters and compare these methods using number of outliers identified. With the help of graph and tables we show the constructions of clusters with outliers.

Fig. 2: Output of K-Means Clustering with cost and number of outliers

Comparison based on number of objects and cost in clusters constructions.

We found that for 50 objects K Mean required 823 costs whereas PAM only required 742 costs, for 100 objects K Mean required 1582 cost whereas PAM only required 1260 costs, for 200 objects K Mean required 3115 cost whereas PAM only required 2472 costs. PAM always required less cost as compared to K Mean. Table 1 shows cost used in cluster construction by K-Mean and PAM and Figure 3 shows comparative graph.

Number of data items	K-Means (Cost)	PAM (Cost)
50	823	742
100	1582	1260
200	3115	2472

Table 1: Number of objects items and clusters cost

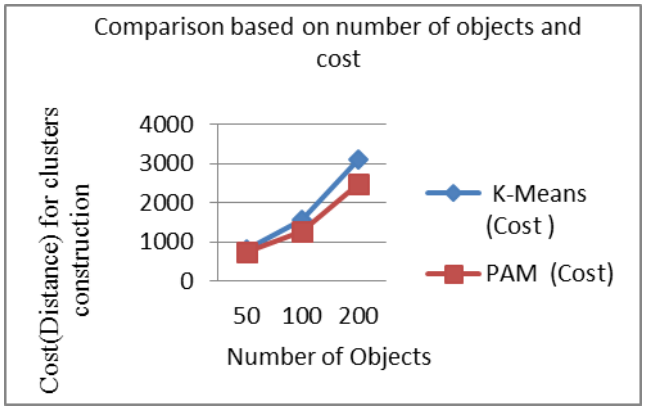

Fig. 3: Number of objects and cost in clusters constructions

X. Conclusion

We compare K Means clustering and PAM (partition around midoids) clustering method based on cost of created clusters and number of outlier identify in both methods. We observed that K mean has more cost as compared to PAM for clusters constructions. In PAM we choose most appropriate center for cluster constructions. Due most appropriate selection of the center PAM handle outlier very efficiently as compared to K Means. We used more 100 students' data to compare the performance of the K Means and PAM clustering method. Clusters are fully dependent on the selection of the initial clusters center. Numbers of center are selected as number for clustered; then distances of all data elements are calculated by Manhattan distance formula. The process is continued until no more changes occur in clusters.

References

[1] Juliano Gaspar "A Systematic Review of Outliers Detection Techniques in Medical Data" 10.5220/0003168705750582 In Proceedings of the International Conference on Health Informatics2011).
[2] Karanjit Singh and Dr. Shuchita Upadhyaya "Outlier Detection: Applications And Techniques" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012 ISSN (Online): 1694-0814 www.IJCSI.org
[3] Jyoti Ranjan Sethi "Study of Distance-Based Outlier Detection Methods" National Institute Of Technology, Rourkela June 2013
[4] Manish Gupta, Jing Gao "Outlier Detection for Temporal Data: A Survey" IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 1, January 2014
[5] Usman Habib, Gerhard Zucker "Outliers Detection Method Using Clustering in Buildings Data" Conference Paper · November 2015 IECON2015-YokohamaNovember 9-12, 2015
[6] Kamaljeet Kaur Atul Garg "Comparative Study of Outlier Detection Algorithms" International Journal of Computer Applications (0975 – 8887) Volume 147 – No. 9, August 2016
[7] Rasim M. Alguliyev, Ramiz M. Aliguliyev "An Anomaly Detection Based on Optimization" I.J. Intelligent Systems and Applications, 2017, 12, 87-96 Published Online December 2017 in MECS

(http://www.mecs-press.org/) DOI: 10.5815/ijisa.2017.12.08

[8] Victoria J. Hodge and Jim Austin "An Evaluation of Classification and Outlier Detection Algorithms Digital Creativity Labs, Department of Computer Science, University of York, UK {victoria.hodge, jim.austin}@york.ac.uk 2 May 2018

[9] Yue Zhao PyOD: A Python Toolbox for Scalable Outlier Detection" Journal of Machine Learning Research 20 (2019) 1-7 Submitted 1/19; Revised 4/19; Published 5/19 arXiv:1901.01588v2 10 Jun 2019

[10] Oyelade, O. J and Oyelade, O. J Application of k-Means Clustering algorithm for prediction of Students' Academic Performance (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, _o. 1, 2010

[11] Madhu Yedla and Srinivasa Rao Pathakota Enhancing K-means Clustering Algorithm with Improved Initial Center Madhu Yedla et al. / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (2) , 2010, 121-125

[12] Juanying Xie Shuai Jiang An Efficient Global K-means Clustering Algorithm Journal Of Computers, Vol. 6, No. 2, February 2011 2011 Academy Publisher doi:10.4304/jcp.6.2.271-279

[13] Bhaskar Mondal and J. Paul Choudhury A Comparative Study on K Means and PAM Algorithm using Physical Characters of Different Varieties of Mango in India International Journal of Computer Applications (0975 – 8887) Volume 78 – No.5, September 2013

[14] Ritu Yadav & Anuradha Sharma Advanced Methods to Improve Performance of K-Means Algorithm: A Review Global Journal of Computer Science and Technology Volume 12 Issue 9 Version 1.0 April 2012 Type: Double Blind Peer Reviewed International Research Journal Publisher: Global Journals Inc. (USA) Online ISSN: 0975-4172 & Print ISSN: 0975-4350

[15] Shraddha Shukla and Naganna S. "A Review ON K-means DATA Clustering Approach" International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 17 (2014), International Research Publications House http://www. irphouse.com

[16] Faisal Bin Al Abid A Novel Approach for PAM Clustering Method International Journal of Computer Applications (0975 – 8887) Volume 86 – No 17, January 2014

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

[18] Manish Gupta, Jing Gao Outlier Detection for Temporal Data: A Survey IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 1, January 2014

[19] Sreevidya S S A Survey on Outlier Detection Methods Sreevidya S S et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 8153-8156

