International Journal of Engineering &Technology, 7 (3.12) (2018) 813 -818

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Outlier Detection using Clustering Techniques

Srividya[1*], S. Mohanavalli[2], N. Sripriya[3], S. Poornima[4]

Department of Information Technology, SSN College of Engineering, Chennai
**Corresponding Author E-mail: [1]srividhyav@ssn.edu.in, [2]mohanas@ssn.edu.in,*
[3]sripriyan@ssn.edu.in, [4]poornimas@ssn.edu.in

Abstract

An outlier is nothing but a pattern that is different compared to the other existing patterns in a particular dataset. In some applications it is very important to understand and identify outliers. Detecting outlier is of major importance in many of the fields like cybersecurity, machine learning, finance, healthcare, etc., A clustering based method is proposed to detect outliers using different algorithms like k means, PAM, Clara, DBScan and LOF on different data sets like breast cancer, heart diseases, multi shaped datasets. This work aims to identify the best suitable method to detect the outliners accurately.

Keywords: *Outliner Detection, Data Mining, K Means ,LOF, CLARA*

1. Introduction

Data mining is extensively used to discover patterns and identify relationship from data. The abnormal patterns in data are called outliers or anomalies or errors or noise or faults or defects. It is very difficult for humans to manually look for anomalies or deviations in large volumes of data. This served as a motivation for building a model to detect outliers in data. The most common causes for outliers are human errors, instrument errors, data processing errors, sampling error, etc., The methods used for outlier detection are discussed as follows.

In the statistical approach, the probability model is used for outlier detection. Different approaches are proposed to detect outliers, and among these the clustering approach is a popular one. Clustering is a tool for outlier analysis. The process of clustering involves grouping of objects with more similarity. The inter similarity between the clusters is very less. Depending on the application, clustering algorithms are chosen. The different types of clustering algorithm available are K Means Clustering, K Medoids Clustering, Density based Clustering, Grid based clustering, etc., The process of clustering is an unsupervised problem as the class labels for the data points are not known before. Outlier detection algorithm first creates normal patterns for given data points and then assigns an outlier score for each data point based on the deviation from the normal pattern.There are three types of outliers based on their composition and relation to the rest of the data [6].

Point outliers (Type I Outliers) If an individual data point is different from other data points, then the data point is a point outlier.

Contextual outliers (Type II Outliers) If an individual data point is dissimilar with respect to context under consideration, then it is called contextual outlier.

Collective outliers (Type III Outliers) If a set or group of data points is different from all other data points, then it is called collective outliers.

Many challenges arise when we detecting the outliers. The way in which outlier is detected changes across different application domains. The constraints and requirements also vary with application. Availability of labelled data or ground truth for training is a challenge in outlier detection. Here, we select and compare some of the available algorithms based on clustering technique to identify which algorithm gives an efficient result. The detection of outliers is very crucial in domains like fraud detection, healthcare, performance analysis, money transactions, etc.,

2. Literature Review

An outlier is also a data point but it deviates so much from the other data points. There are different causes that leads to the presence of outliers. This notion is formalized by Knorrand Ng in the definition of outliers [18]. In this paper, the LOF (local outlier factor) is used to detect outliers [16]. Technique used in outlier detection is based on the application domain. Each domain has a different definition for outlier detection. The problem with detecting Type II is discussed [6]. In his paper, Dr. T. Christopher discusses about comparison of algorithms in outlier detection. It is observed that the accuracy is better when applying the CURE with CLARANS than CURE with K-Means clustering [15]. When the data stream is used for outlier analysis KORM (K-Median Outlier Miner) K-mediod gives better performance than k-means [2]. In this paper, built in healthcare dataset like ESOPS, diabetes and Kos tecki Dillon of R is used for outlier analysis. Cluster-based algorithms are found to give better accuracy than distance based outlier detection methods [4]. Hierarchical clustering algorithm is used for foreign trade transaction dataset to detect outlier. This method uses the size of the resulting clusters for identifying groups of observations that are outliers [5]. It uses probability model and dynamic link optimization algorithm for finding outliers, as it gives better performance in said dataset. Outliers are confirmed by Dynamic Threshold Optimization algorithm [11]. A new approach is performed in to detect outlier by using PAM; outliers are detected in two phases. PAM algorithm is used to produce set of clusters. In the first phase, clusters are checked for number of data points contained. If it contains less than average number of datapoints then the whole set of datapoints is considered as an outlier. In second phase absolute distance between mediod (ADMP) and other point is calculated; each cluster has some threshold value, and if ADMP is greater than threshold then that object is an outlier [12]. According to Alberto M. C. Souzaa, Jos´e R. A. Amazonas used new technology like the Hadoop framework along with IOT LinkSmart middleware to detect the outlier. It is proposed to give flexibility and scalability [3]. CURE algorithm is efficient when compared to other clustering algorithm like K-means, CLARA, and CLARANS. When handling high dimensional data, it is important to select the clustering algorithm based on the dataset type. In this research work many unsupervised clustering algorithms are studied for outlier detection technique [14]. Statistical method is totally dependent on the probability model. Many statistical methods are there to detect the outliers like Test, Smart sifter, Regression analysis, Replicated neural network [1]. Cluster based OutlieRMinera (CORM) is presented by Elahi. This clustering-based approach for outlier detection is based on K-means clustering algorithm. The consideration of all attributes results in poor performance [17]. The Mahalanobis distance measure can also be used for detection of outliers.

The concept of correlation is used to identify patterns. It is scale-invariant and differs from Euclidean distance. [13]. A novel method has been proposed using neighbourhood rank difference for the detection of outliers. The results of experiments are recorded for real and synthetic datasets with different dimensions [7]. This paper proposes a framework for detecting outliers in evolving data streams. The authors have discussed the effects of assigning weights to attributes. [8]. In this work [10], the authors discuss the various techniques for outlier detection of temporal data.

3. Empirical Study

In study of outlier detection has been implementedusing the algorithm of clustering, which is based on density based clustering and distance based clustering. Then comparison of which algorithm detects outliers the best. These algorithms are clearly explained below, with their pseudo code and the implementation steps. Clustering is the process of grouping related objects or more similar objects. Similarity among the clusters is very less. Clustering is an unsupervised technique for outlier detection; it gives better results among other techniques.

Table 1: Type of Clustering Algorithms

CLUSTERING METHOD	ALGORITHM
Distance based clustering	K Means
Distance based clustering	PAM
Distance based clustering	CLARA
Density based clustering	DBSCAN
Density based clustering	LOF

K Means Algorithm

The k-means clustering algorithm partitions a a given data set into a fixed number k of clusters. For cluster initially a centroid is chosen. A centroid is a data point at the center of a cluster. This algorithm fails identify clusters of arbitary shape, it is difficult to ascertain if the obtained k is indeed correct, and also does not handle clusters with different densities.

K- Medoids Algorithm

K-Medoids is similar to K-Means Algorithm but it uses medoids to form a cluster. There are two types of K-Medoid algorithms - PAM (partition around medoids) and CLARA. The algorithm chooses actual datapoints as centroids. The datapoints that are closer to the centroid forms a cluster and the others being a point that belongs to some other group or simply an outlier. PAM performs better than k-means even with noisy data and works efficiently for small data sets.

CLARA (Clustering Large Applications)

CLARA is introduced to overcome the drawbacks faced by PAM. It offers the advantage of working with smaller samples of a large dataset. The PAM algorithm is used to choose the medoid. The choice of sample directly affects the efficiency of the algorithm.

DBSCAN

DBSCAN is a density based clustering algorithm. The datapoints that are identical and also closely packed are grouped and marking others as outliers points that lie alone in low-density regions. The advantage is no prior knowledge of the number of clusters is required. It is able to identify noise data in a better manner and is able to detect arbitrarily shaped clusters with different sizes. The algorithm does not scale well for high dimensional data and varying densities.

LOF (Local Outlier Factor)

In local outlier, locality is the distance that is used to estimate the density. A region with more number of closely packed datapoints is termed as high density. The number of datapoints that discriminates high and low density is decided by comparing densities in neighbourhood groups. The use of local density approach makes it easy to identify outliers using LOF. The advantage of LOF is that is clearly able differentiate and inlier and an outlier. This method of detecting outliers can also be generalized to work effectively across different datasets.

4. Experimental Results

The outlier detection project did some implementation by using clustering algorithms on different datasets like healthcare and other default datasets. By using these implementations, the best algorithms for detecting outliers were concluded. This implementation uses many datasets. The below implementation all are done in the R language by using R software.

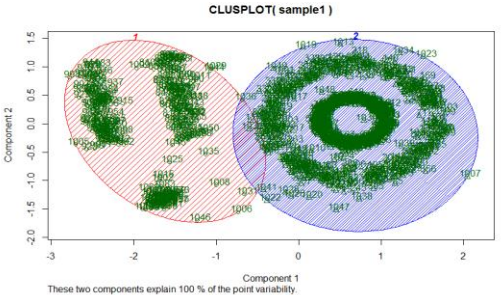

Fig. 1: K-Means for Multi shapes dataset

Description of Datasets

Some datasets are used for implementing algorithms. The below explanation going to explain about the dataset which was used in this implementation and we used numeric data for easy implementation. Breast cancer dataset get from UCI repository. This is a numeric dataset and it has 198 observation and 14 variables. Then this dataset form a cluster based on the variable called outcome. That outcome variable has two values called Recurrence and non-Recurrence

Multi shapes dataset get from the package FACTOEXTRA which is default dataset in R Language. This is a numeric dataset and it has 1100 observation and 3 variables. This has several shapes and that is defined by angle value.

Student dataset is taken as a synthetic dataset. This is also a numeric dataset and it has 36 observation and 5 variables. Then the cluster is formed based on the variable called result which has two variables called pass and fail.

Hungarian heart disease dataset is used and taken from UCI repository. It is a numeric dataset and it has 14 attributes like cholesterol, blood pressure, pain type, ECG graph etc. This dataset has 294 observations outlier is identified based on the difference in above attribute values.

K-MEANS Algorithm

K -Means algorithm is a method which is based on the distance based clustering. This method uses the random center point and the mean value is taken to form the clustering. Then the point which is more distant from the cluster they are all consider as a outliers. This outlier points are check by outlier function in R language.

K-Means for Multishapes dataset

This multishapes dataset get from **factoextrapackage.** It is a default dataset. From this implementation we get two clusters and we can findoutliers visually which the points are all in long distance from clusters. This algorithm forms a cluster based on a shapes of the angle value which is in dataset. K means algorithm is best suited for forming clusters.

K-Means implementation for breast cancer

This dataset is obtained from the UCI repository. The plot has two clusters which is form based on the attribute outcome.

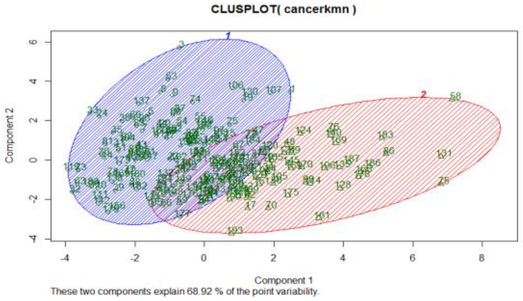

Fig. 2: K-Means for Breast Cancer dataset

That has a value One for Recurrence and the other one for Non recurrence persons. The plot represents the cluster of dataset by using **Principal Component Analysis(PCA)** method.

K-means Implementation for Synthetic Dataset

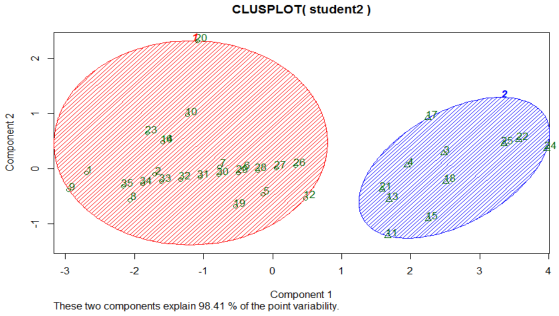

Fig. 3: k-means for synthetic dataset

Synthetic dataset is a student dataset, some artificial outlier are included to conclude which algorithm is best. Clusters are formed based on the pass and fail criteria. We insert some random values into data to detect the efficiency of the algorithm for detecting outliers .It identifies 50% of outliers which we inserted in the dataset.

K-Means implementation for heart disease dataset

This dataset has significant attribute like cholesterol, blood pressure, blood sugar etc are based on the variance in this attribute. Based on the size of vessel it gives the vessel narrowing size. It forms a cluster which is less than 50% narrowed and which is more than 50% narrowed.

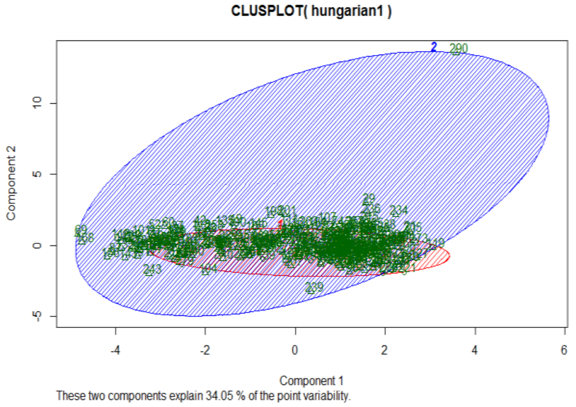

Fig. 4: k-means for heart disease

DBSCAN algorithm implementation

In DBSCAN algorithm the clustering form is based on reachability point. That point is based on local density between the each points.

DBSCAN for Multi shapes:

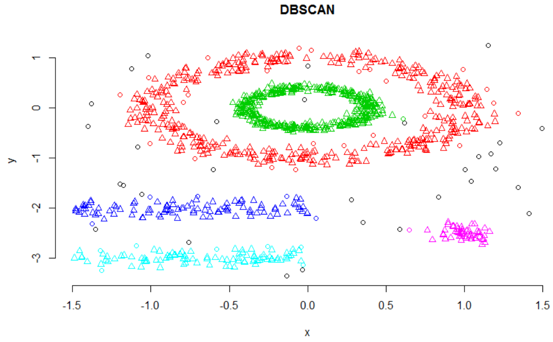

Fig. 5: DBSCAN for Multi shapes dataset

The above plot has five clusters. It's done based on the minimum point and an eps value which is given in the algorithm while implementation. If the density is low then those point all are consider as an outliers are cannot accepted by clusters. So they are outside the cluster and they are represented in the above plot by black colour point.

DBSCAN for heart disease

Outliers are denoted in red colour triangle shape which is deviated from whole dataset. Clusters are formed based on the density. Variable for x and y axis chosen by PCA method .It analysis all the component and select which has most significant variance.

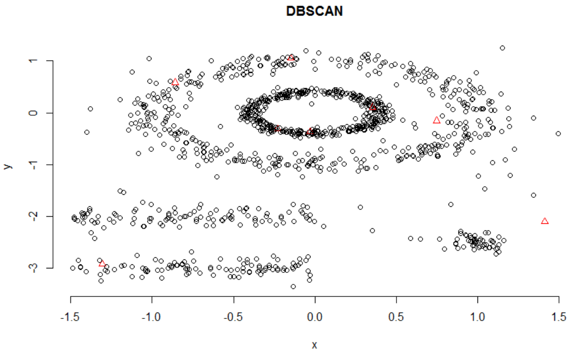

Fig. 6: DBScan for heart disease

PAM for multishapes

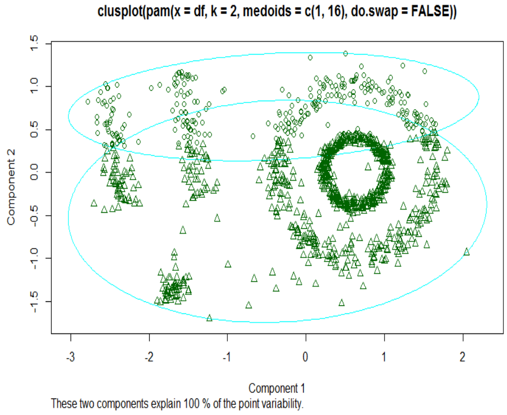

Fig. 7: PAM for Multi shapes dataset

The algorithm is intended to find a sequence of objects called medoids that are centrally located in clusters and group the point which is closest to the medoid. This is one of the types of K-Medoid algorithm. In this method this algorithm took all the points to perform the operation.

PAM implementation for breast cancer data

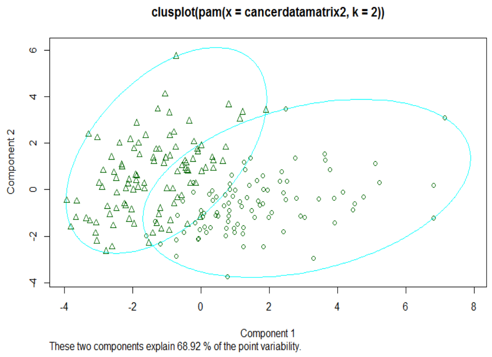

Fig. 8: PAM for breast cancer data

Based on the recurrence and non-recurrence variables, clusters are formed. Cluster is formed based on the median instead of mean value. It is robust than k-means algorithm when outliers are present in dataset.

PAM for synthetic dataset

PAM algorithm for synthetic dataset is performed based on the result which is pass or fail, we created student dataset as a Synthetic dataset for implementation.

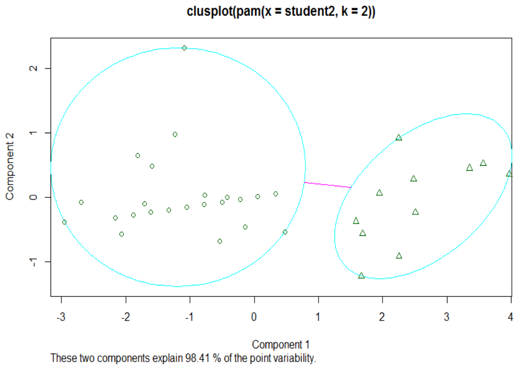

Fig. 9: PAM for synthetic dataset

PAM algorithm for synthetic dataset is performed based on the result which is pass or fail; we created student dataset as a synthetic dataset for implementation.

PAM for heart disease dataset

In heart disease the cluster is formed based on cholesterol and the outliers also show in the below diagram based on the cholesterol.

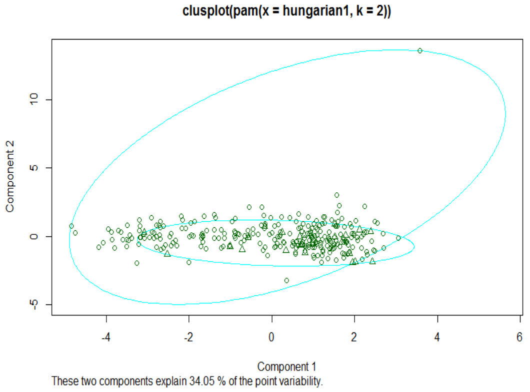

Fig. 10: PAM for heart disease dataset

CLARA for multishapes

In this method this algorithm took many sample data points from whole dataset and perform PAM algorithm for each sample and produce the result based on all the sample implementation.

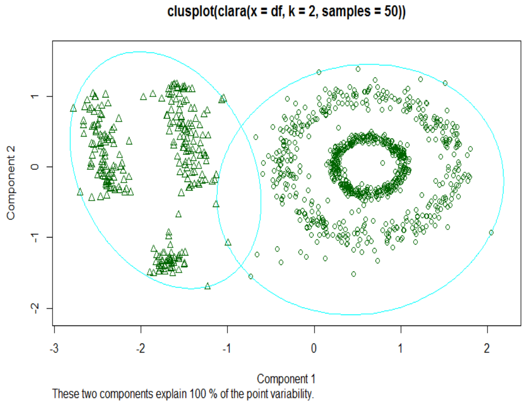

Fig. 11: CLARA for Multi shapes dataset

Large numbers of samples are used for CLARA implementation. It is like large dataset is divided into samples and then PAM is applied. Based on the angle values clusters are formed.

CLARA for breast cancer dataset

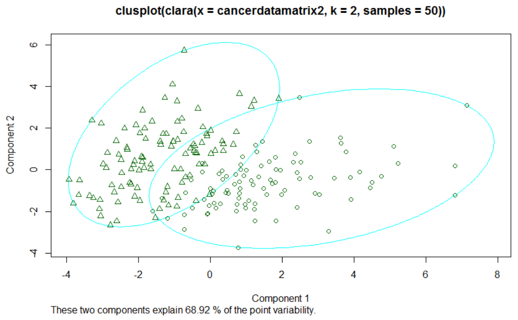

Fig. 12: CLARA for breast cancer dataset

Based on the recurrence and non-recurrence values, clusters are formed 50 samples are used to implement clustering algorithm.

CLARA for synthetic dataset

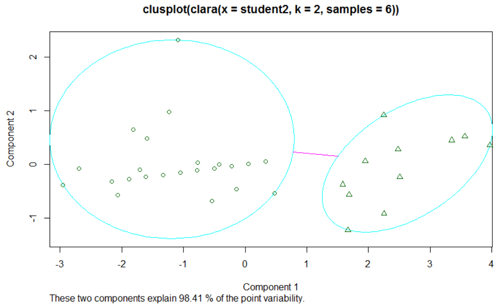

Fig. 13: CLARA for synthetic dataset

CLARA for heart disease

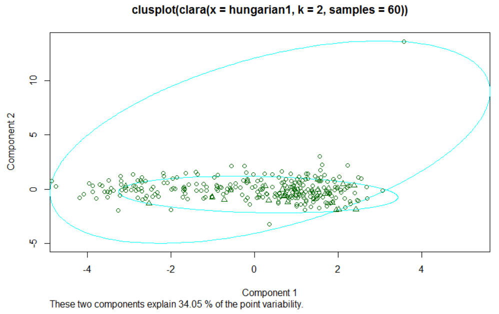

Fig. 14: CLARA for heart disease

CLARA uses samples of data to form a cluster it is useful when large dataset is used.

LOF Algorithm Implementation

The local outlier factor (LOF) is a measure of deviation of a datapoint. This measure can be used to decide if the datapoint is an outlier or not. The LOF considers the density of the neighbours to decide whether a point is an outlier. The arrow mark indicates the dimensions of dataset. This dataset has a more than two dimensions. So this plot implementation uses a PCA method. This is more simple and efficient for detecting outliers.

LOF for Multishapes

This algorithm is about the implementation of **c**lusters formed based on local density between each point in a dataset. The function lofactor (data ,k) is used to calculate the local outlier factor ,where k is number of neighbors used in calculation of Local Outlier Factor score. Based on the Local Outlier Factor score, which has low density is considered as an outlier.

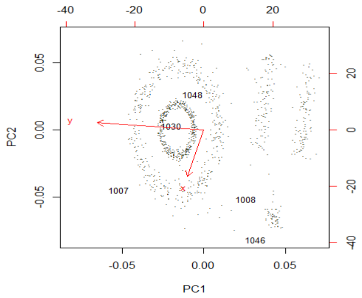

Fig. 15: LOF for Multi shapes dataset

LOF implementation for breast cancer dataset

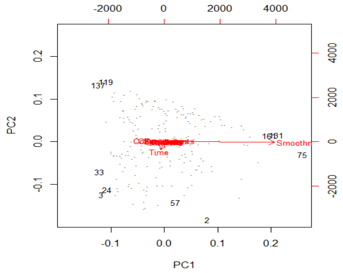

Fig. 16: LOF for Breast cancer dataset

This implementation forms a clustering based on the attribute called outcome which has a value recurrence and non recurrence of local reachability. Calculate the local score and find the local density. Then the low local density is considered as an outlier point.

LOF for synthetic dataset

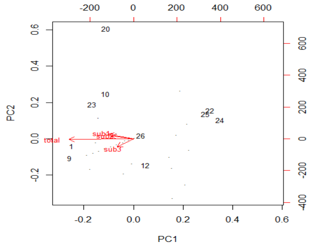

Fig. 17: LOF for synthetic dataset

Plot denotes which attributes are varied based on that outliers are identified. Some random values are inserted to define the efficiency of LOF algorithm for detecting outlier .It identified 75% of artificial outliers which we inserted. The variable that is denoted by the long arrow has lot of difference.

LOF for heart disease dataset

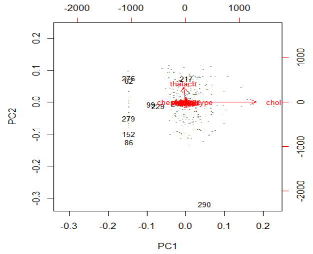

Fig. 18: LOF for heart disease dataset

This algorithm form a cluster based on the attribute called colastral. Then it calculates the local score which is nothing but a local density between each point. Then based on that local density the outliers are shown in the figure. This is efficient method for detecting outliers. In this case it detects 75% of the outliers.

5. Future Work and Conclusion

The goal of comparing the various algorithms in the paper is to improve the quality of detecting the outlier by using clustering. In this paper, the LOF algorithm is concluded to be the best algorithm for detecting outliers as LOF gives more importance to detecting local outliers than the other methods. The main reason for the better performance of LOF is that it uses the single parameter to implement the algorithm. When we consider the global outlier it gives a lower performance for detecting outliers, but in the local outlier, each and every point in the dataset is analysed to produce the best outlier detection performance. So we can find the outlier by using LOF efficiently. From our results of implementation it found that 75% of the outliers are found. This efficiency can vary based on the applications and datasets available. Nowadays, the outlier detection methods are used in many applications like fraud detection, Healthcare data, Student performance and Sports data. As far as the study is concerned, we have worked on a stored data set which contains the outliers. In future, we will work on data streams, because outlier detection on data streams has many challenges. In data streams, values change continuously. Hence the detection of outlier becomes difficult. Our future work will be focusing on these algorithms to detect outliers efficiently in the data stream.

References

[1] Petrovskiy, M. I. "Outlier detection algorithms in data mining systems." Programming and Computer Software 29.4 (2003): 228-237.

[2] Dhaliwal, Parneeta, M. P. S. Bhatia, and Priti Bansal. "A cluster-based approach for outlier detection in dynamic data streams (KORM: k-median OutlieR miner)." arXiv preprint arXiv:1002. 4003(2010).

[3] Souza, Alberto MC, and Joseé RA Amazonas. "An outlier detect algorithm using big data processing and internet of things architecture." Procedia Computer Science 52 (2015): 1010-1015.

[4] Christy, A., G. Meera Gandhi, and S. Vaithyasubramanian. "Cluster Based Outlier Detection Algorithm for Healthcare Data." Procedia Computer Science 50 (2015): 209-215.

[5] Loureiro, Antonio, Luis Torgo, and Carlos Soares. "Outlier detection using clustering methods: a data cleaning application." Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. Bonn, Germany. 2004.

[6] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Outlier detection: A survey." ACM Computing Surveys (2007).

[7] Bhattacharya, Gautam, Koushik Ghosh, and Ananda S. Chowdhury. "Outlier detection using neighborhood rank difference." Pattern Recognition Letters 60 (2015): 24-31.

[8] Toshniwal, Durga. "A framework for outlier detection in evolving data streams by weighting attributes in clustering." Procedia Technology 6 (2012): 214-222.

[9] Cao, Lei, Qingyang Wang, and Elke A. Rundensteiner. "Interactive outlier exploration in big data streams." Proceedings of the VLDB Endowment 7.13 (2014): 1621-1624.

[10] Gupta, Manish, et al. "Outlier detection for temporal data: A survey." IEEE Transactions on Knowledge and Data Engineering26.9 (2014): 2250-2267.

[11] SREEVIDYA, SS. "Detection of Outliers in Data Stream Using Clustering Method." International Journal of Science, Engineering and Technology Research (IJSETR)/2015/2278-7798 4 (2015).

[12] Kumar, Vijay, Sunil Kumar, and Ajay Kumar Singh. "Outlier Detection: A Clustering-Based Approach." International Journal of Science and Modern Engineering (IJISME), ISSN (2013): 2319-6386.

[13] Jayakumar, G. D. S., and Bejoy John Thomas. "A new procedure of clustering based on multivariate outlier detection." Journal of Data Science 11.1 (2013): 69-84.

[14] Papadimitriou, Spiros, et al. "Loci: Fast outlier detection using the local correlation integral." Data Engineering, 2003. Proceedings. 19th International Conference on. IEEE, 2003.

[15] Christopher, T., and T. Divya. "A Study of Clustering Based Algorithm for Outlier Detection in Data streams." Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications. 2015. National Conference on Advanced Networking and Applications, 27th March 2015.

[16] Breunig, Markus M., et al. "LOF: identifying density-based local outliers." ACM sigmod record. Vol. 29. No. 2. ACM, 2000.

[17] Elahi, Manzoor, et al. "Efficient clustering-based outlier detection algorithm for dynamic data stream." Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on. Vol. 5. IEEE, 2008.

[18] Knox, Edwin M., and Raymond T. Ng. "Algorithms for mining distance based outliers in large datasets." Proceedings of the International Conference on Very Large Data Bases. Citeseer, 1998.

[19] Singh, Janpreet, and Shruti Aggarwal. "Survey on outlier detection in data mining." International Journal of Computer Applications67.19 (2013).

[20] Pachgade, Ms SD, and Ms SS Dhande. "Outlier detection over data set using cluster-based and distance-based approach. "International Journal of Advanced Research in Computer Science and Software Engineering 2.6 (2012).

[21] Pamula, Rajendra, Jatindra Kumar Deka, and Sukumar Nandi. "An outlier detection method based on clustering." Emerging Applications of Information Technology (EAIT), 2011 Second International Conference on. IEEE, 2011.