International Journal of Latest Trends in Engineering and Technology
Vol.(7)Issue(3), pp. 319-325
DOI: http://dx.doi.org/10.21172/1.73.544
e-ISSN:2278-621X

An Experimental Analysis of Outliers Detection on Static Exaustive Datasets.

Raghav M Purankar[1] and Prof. Pragati Patil[2]

Abstract-Detecting Outlier and clustering methodologies are an important branch of data mining, by combining the two technologies can improve the data mining significance. Exhaustive data sets by experimental results ensures that the algorithm will improve the efficiency of outlier detection. The outliers may be instances of error or indicate different behavior. The task of outlier detection primarily focused on identifying such outliers so as to improve the data analysis and then find out only interesting and useful information about unusual events within number of application domains. Finding outliers in a group of patterns is a very well-known issue in the data mining. The principle of outlier detection depend on the threshold value. Threshold is generally provided by user. In proposed approach, two methods cluster based approach and distance based approach are applied individually over static data sets to efficiently find the outlier from the data set. The exhaustive datasets can be downloaded from the UCI machine learning repositories. The experimental results of real exhaustive dataset demonstrate that proposed method takes lesser cost in computation and gives better performance than the traditional methods.

Keywords –Cluster based Approach, Data Mining, Distance based approach, Outlier Detection, UCI Repository.

I. INTRODUCTION

Data mining, in general, focuses on the finding the non-trivial, hidden and useful interesting information from various types of data with the advancement of Information Technologies. Data streams are ubiquitous. These can be found in many application domains from financial transactions to be done online to medical domain and space research centers, where satellites are continuously generates data streams. Clustering, Classification and Association has vital correlation with data mining[1]. In recent years, existing database querying methodologies are not sufficient to extract useful information, and hence researchers nowadays are primarily aiming towards development of new techniques to meet the improved requirements. Outlier detection is an important and major research issue that aims to find objects. that are considerably different, abnormal and inconsistent in the database. Efficient and effective detection of outliers minimizes the risk of making poor decisions based on erroneous data, and helps in identifying, preventing from the effects of malicious or faulty behavior of data. One of the important factor in data mining is increase in dimensionality of data gives rise to a number of computational challenges

[1] *Department of Computer Science and Engineering AGPCE, Nagpur University, Nagpur, Maharashtra, India*
[2] *Department of Computer Science and Engineering AGPCE, Nagpur University, Nagpur, Maharashtra, India*

and issues. Nowadays, it is observed that most of the research activity starts with the explosion of collected data and to convert it in the format of data stream.

An outlier detection in infinite and massive streaming data is one of the active research area of data mining that primarily aims to finding object which have dissimilar behavioral properties than normal object. An outlier is an object that is significantly different or inconsistent to rest of the data object where as fraud identification, generated web logs, and available online documents are few application areas of outlier detection in data stream domain. There are numerous algorithms for outlier detection of stored static data sets which are based on a variety of approaches like nearest neighbor based, distance based detection and clustering based outlier detection[2]. The evolution of streaming data led to the change in the properties and features of the data streams such as dimensionality, anomalous features of object of a point which is an outlier in previous stage may become an inlier in the further stages. Accurate and efficient removal of outliers may greatly improve the performance of statistical data mining techniques[3,4]. Data cleaning is the process of finding and removing outliers as a preprocessing step using clustering methodology.

II. METHODOLOGY

A. Datasets –

In order to compare the clustering of data stream for finding outliers, datasets are taken from UCI machine learning repository. In this case we manually added few no. of outliers into dataset file and tried to compare the result of finding the outlier by running the K-means algorithm over it[5]. Then find out the efficiency of algorithm for detecting outlier from the dataset. In this paper we perform experiments on Iris datasets and tried to find out the anomalies. We have forcefully inserted the outliers at specific location and run the algorithm over the data file, algorithm gives maximum efficiency.

B. Clustering –

The Data clustering is an important but quite difficult problem. Clustering methodology requires the definition of a similar properties between patterns, which is not easy to specify without having any prior knowledge about cluster shapes[8]. Clustering is not a new concept but data clustering specifically for outlier detection is a recent discipline under continuous development. In data streams the clustering is one of the sub-process which is used to group the similar kind of objects as well as it is used to detect the anomalous object efficiently and also clustering is one of the unsupervised method in data streams[6]. Clustering-based methods assume that the normal data objects belong to large clusters, whereas outliers belong to small or different clusters or may not belong to any of the clusters[14]. Clustering-based methods find out outliers by detecting the similar kind of properties between objects and clusters.

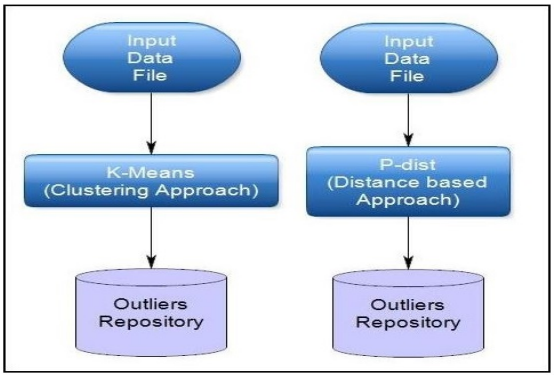

Figure 1. Cluster Approach and Distance Approach.

C. Outlier Detection –

The Outlier detection is a very much important step in many data-mining applications area. It refers to the process of finding patterns in data that do not have expected normal behavior or anomalous behavior. It is further advancing by the fact that in many cases outliers have to be find out from a large volume of data growing at an infinite rate. Traditional algorithms for outlier detection cannot be so efficient and effective to data stream, since the streaming data is potentially infinite and evolving in nature. With the advancement of the medical data increasing continuously, the process of determining outliers becomes more complicated and tedious.

III. PROPOSED APPROACH

The proposed work uses two individual approaches such as clustering and distance based approach to formulate the cluster and Outlier Detection using static data sets downloaded from UCI machine repository[23]. The aim of proposed work is to predict outlier detection on numeric data by using existing cluster based and distance approach and we are trying to improve the efficiency of outlier detection using two individual approaches over the same static data sets[7]. First of all static data sets are filtered or preprocessed so that we are able to perform the analysis only on desired attributes. Preprocessed data then given as input to the K-means algorithm for clustering and then apply distance based algorithm P-Dist on the same data and find out the anomalous data point. As the number of outlier are not sufficient because above algorithms are not so effective, thus we carry out the operation using two consecutive algorithm by little modification in outlier detection approach. Proposed system gives the idea about outlier detection by windowing technique. Thus Proposed methodology gives better performance for outlier detection as compared to traditional individual approach.

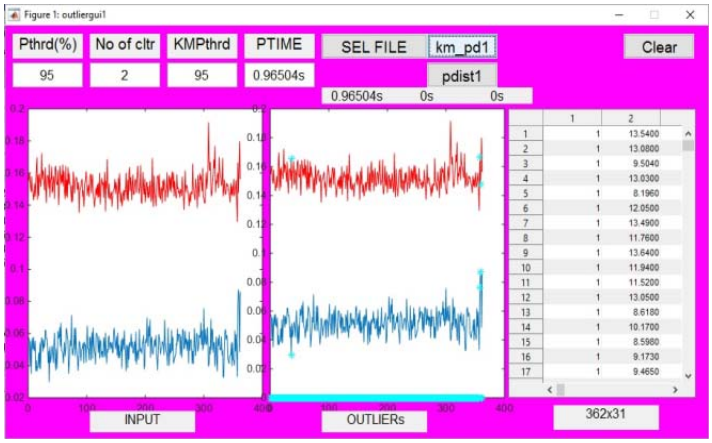

Figure 2. K-means Outlier detection approach algorithm.

Above figure shows the GUI for K-means algorithm to operate over the static Iris data set downloaded from UCI machine repository. as in the above clustering approach K-means algorithm does not give greater accuracy of outlier detection, so distance based approach is used by P-dist algorithm.

IV. EXPERIMENTAL RESULTS

In this section we have made comparative analysis of outlier detection method over different datasets. In outlier detection method, we used K-means and P-dist algorithm and define threshold value for each algorithm while executing. We select dataset file, Added some known outliers into dataset file and perform the K-means algorithm with setting the threshold value , The resulting output shows the number of outliers and index locations where the outliers are Added. Similarly the process is followed for P-dist algorithm and compares the results for both the algorithms. If the output results are suffices, then we can verify that both the algorithms are running the appropriate task for outlier detection.

Table -1 Execution time and Accuracy of Outliers Detection using K-means.

K Means Threshold (90%)			
Number of Outliers Added	**Number of Cluster**	**Execution Time**	**Approximate Accuracy**
5	**2**	**0.217**	**40**
6	**2**	**0.176**	**33.33**
14	**2**	**0.21**	**28.57**
27	**2**	**0.19**	**7.4**
36	**2**	**0.216**	**2.77**

We have perform the experimental analysis using K-means algorithm on the Iris data set downloaded from the UCI machine repository and calculate the running time and approximate value of outlier detection[24].In Iris data set, there are total 150 instances. These instances are divided into three classes of 50 instances in each class. We have consider some of the attributes of Iris data set as sepal length, sepal width, petal length and petal width. The threshold we kept constant throughout the analysis taken i.e.90%. From the table it is shown that Efficiency for outlier detection goes on decreasing as the size dataset increases. K-means algorithm is basically used for clustering purposed[13], so we may not get improved efficiency of outlier detection. For this reason we tried to implement pair wise distance approach to maximize the efficiency of outlier detection to certain comparative level.

Similarly, for P-dist algorithm we have calculated the operational efficiency at the constant threshold of 95%. Here while running the P-Dist algorithm execution time goes decreasing and the efficiency is improving but we can achieve maximum 80% of accuracy after that the efficiency remains constant. From the above analysis it is noted that the running time for the K-means algorithm is more than the time required for P-dist algorithm.

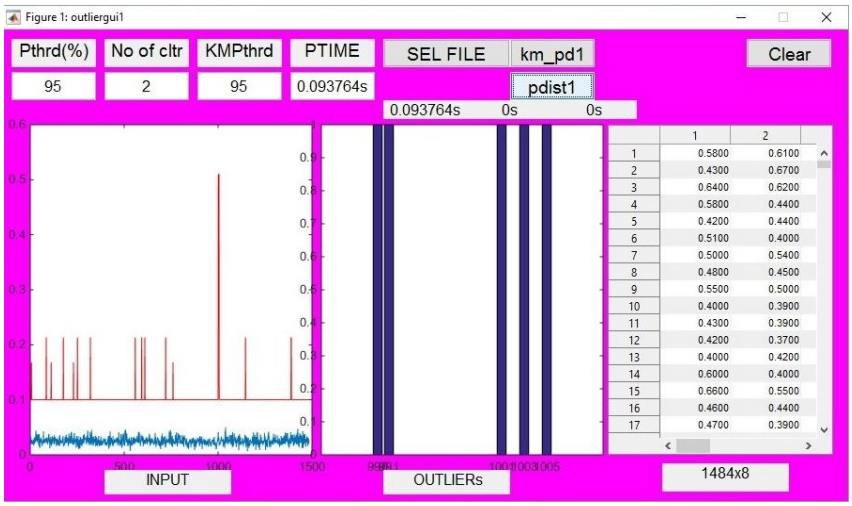

Figure 3. P-Dist Outlier detection approach algorithm.

Above figure shows the experimental screenshot when we run p-dist algorithm over the input data set by using threshold value at 95%. Middle panel shows the Bar graph at a particular location where outliers are present in the data file.

Table -1 Execution time and Accuracy of Outliers Detection using P-Dist.

Pairwise Distance Threshold (95%)			
Number of Outliers Added	**Number of Cluster**	**Execution Time**	**Approximate Accuracy**
5	**2**	**0.058**	**72**
6	**2**	**0.0091**	**75**
14	**2**	**0.0072**	**76**
27	**2**	**0.041**	**80**
36	**2**	**0.038**	**80**

As indicated and drawn from the experimental results that the accuracy of outlier detection is comparatively greater in case of distance based approach than clustering approach, and the accuracy will remains constant after certain value. The entire result sets are written into the excel file which is stored at the same location from where we have taken the input data file.

V. CONCLUSION

In this paper we proposed basically two individual approaches for detecting the outlier in static exhaustive data sets. The method applied both clustering and distance based approach consecutively using K-means and P-dist algorithm for detecting outliers from Iris data set.

Outlier detection depending upon the relevance of each data elements. The two individual approaches are used back to back in order to improve the accuracy of outlier detection.

The proposed method is continuous and more effective in nature for accepting the challenges of large exhaustive static data sets. In this proposed method, we are trying to improve the accuracy of outliers detection since K-means does not provide higher accuracy when operated individually, so next we used P-dist approach so that we get better accuracy of outlier detection. In future scope we will try to take our approach to next higher level like hybrid approach for dynamic data stream where we will try to use the combined methodologies for dynamic data stream.

REFERENCES

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine Volume 17 Number 3 (1996).

[2] Ms. S. D. Pachgade, Ms. S. S. Dhande, "Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach", International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 6, June 2012.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection:A Survey," ACM Computing Surveys, vol. 41, no. 3, pp.15:1–15:58, 2009.

[4] Hans-Peter Kriegel, Peer Kröger, Arthur Zimek, "Outlier Detection Techniques", The 2010 SIAM International Conference on Data Mining, Tutorial Notes: SIAM SDM 2010.

[5] Dharmendra S. Modha and W. Scott Spangler, "FeatureWeighting in k-meanss Clustering",2002 Kluwer Academic Publishers. Printed in the Netherlands, Machine Learning, Vol. 47, 2002.

[6] Reddy M.V. Jagannatha and B. Kavitha, "Clustering the Mixed Numerical and Categorical Dataset using Similarity Weight and Filter Method", International Journal of Database Theory and Application Vol. 5, No. 1, March, 2012.

[7] Knorr E.M., Ng R.T., Tucakov V., "Distance based method: algorithm and applications", International Journal of Soft Computing and Engineering, Volume. 3, Issue 6, January 2014.

[8] Rajendra Pamula,"Outlier detection method based on clustering", Emerging Applications of Information Technology (EAIT), Second International Conference, February 2011.

[9] Sridhar Ramaswamy, "Efficient algorithms For mining outliers from large datasets", SIGMOD '00 Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Volume 29 Issue 2, June 2000.

[10] K. Das and J.G. Schneider, "Detecting anomalous records in categorical datasets, " in Proc. of the Second Int. Conf. on Knowledge Discovery and Data Mining, San Jose, California, USA, pp.220-229, August 2007.

[11] Ramaswamy S., Rastogi R., Kyuseok S.:Efficient Algorithms for Mining Outliers from Large Data Sets,Proc. ACM SIDMOD Int.Conf. on Management of Data, 2000.

[12] F.Angiulli, S. Basta, and C. Pizzuti. Distance-based detection and prediction of outliers. IEEE Transactions on Knowledge and Data Engineering, 18:145–160, 2006.

[13] S. Guha, R. Rastogi, and K. Shim. CURE: An efficient clustering algorithm for large databases. SIGMOD Rec., 27(2):73–84, 1998.

[14] Hanwen Yu, Zhenfang Li, and Zheng Bao, Residues Cluster-Based Segmentation and Outlier-Detection Method for Large-ScalePhase Unwrapping. IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 20, NO. 10, OCTOBER 2011.

[15] C. C. Aggarwal, Outlier Analysis. Springer, 2013.

[16] C. Aggarwal and K. Subbian, "Event Detection in Social Streams," in Proc. of the 12th SIAM Intl. Conf. on Data Mining (SDM), 2012, pp. 624– 635.

[17] M. Gupta, J. Gao, and J. Han, "Community Distribution Outlier Detection in Heterogeneous Information Networks," in Proc. of the 2013 European Conf. on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), 2013, pp.557–573.

[18] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier Detection for Temporal Data," in Proc. of the 13th SIAM Intl. Conf. on Data Mining (SDM), 2013.

[19] M. Gupta, A. B. Sharma, H. Chen, and G. Jiang, "Context-Aware Time Series Anomaly Detection for Complex Systems," in Proc. of the SDM Workshop on Data Mining for Service and Maintenance, 2013.

[20] Dr. S. Vijayarani and Ms. P. Jothi, Detecting Outliers in Data streams using Clustering Algorithms, International Journal of Innovative Research in Computer and Communication Engineering, Volume. 2, Issue 8, October 2013.

[21] D.Joice, K. Lakshmi and K. Thilagam, Comparison Of Cluster Based Algorithms For Outlier Detection In High Dimensional Dataset, Karpagam Journal of computer science, Volume 8, issue 3, April 2014.

[22] Dr. S. Vijayarani and Ms. P. Jothi, Partitioning Clustering Algorithms For Data Stream Outlier Detection, International Journal of Innovative Research in Computer and Communication Engineering, Volume. 2, Issue 4, April 2014.

[23] http://archive.ics.uci.edu/ml/

[24] http://archive.ics.uci.edu/ml/machine-learning-databases/iris.