Comparative Study of Outlier Detection Algorithms

Kamaljeet Kaur
Research Scholar
MMICT&BM, MM University
Mullana, Ambala

Atul Garg
Associate Professor
MMICT&BM, MM University
Mullana, Ambala

ABSTRACT
As the dimension of the data is increasing day by day, outlier detection is emerging as one of the active area of research. Finding of the outliers from large data sets is the main problem. Outlier is considered as the pattern that is different from the rest of the patterns present in the data set. The detection of the outlier in the data set is an important process as it helps in acquiring the useful information that further helps in the data analysis. Various algorithms have been proposed till date for the detection of the outliers. This paper 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.

Keywords
Outlier Detection, Statistical Outlier Detection, Density based, Clustering, Classification.

1. INTRODUCTION
With the rise in the various technologies, the amount of data along with its dimensions and complexity is growing so rapidly that for the great amount of information various automated analysis techniques are required. There are various reasons that can induce outlier in the data; some of them are malicious activities like credit card fraud, cyber activity, breakdown of system, mechanical faults, changes in system behavior etc. Outliers are the observations whose actual value is different than the rest of the observed value of the data [20]. These observations are deviated due to the errors present in data collection process. This can cause problem at the time of analyzing the result. Removal of outliers from the data sets helps in retrieving the useful and meaningful knowledge. It also improves the data analysis for further research within various application domains [14].Various conventional methods have been used in data mining community for detection of outliers [7]. These conservative techniques are more suitable on static data sets rather than handling dynamic nature of data.

With the increase in the dimensionality of the data, the techniques applied on the streaming data or large data sets cannot provide an efficient result and also takes high computation time [17]. Construction of the model is required that perfectly represents the data for the effective outlier detection. Over the years, many outlier detection techniques have been developed for anomaly detection and developing models for outliers. Various problems arise due to this increased size of data like data redundancy, missing data etc [32]. The outlier detection from the data helps to acquire the useful knowledge that will help in data analysis. This can be used for different applications in different domains that will give the efficient results. The rise in the data dimensionality is considered as the major problem in the process of data mining [28].

The Figure-1 defines the process of outlier detection. Initially a network is created. Network is basically the area or the region in which the dataset is present and from this dataset the outliers are detected. In next step the data is taken from the network and is converted into log files. By converting the accumulated data into log files the outlier can be easily detected. Then the points that are different from the rest of the points of the dataset are detected

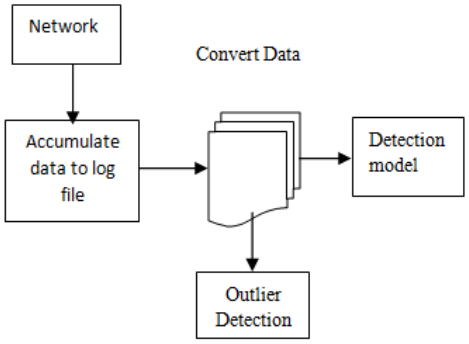

Figure-1 General Process of Detecting Outliers

The main objective of detecting outliers is to retrieve the objects from the large datasets that have different behavior than the normal object present in the data [16]. Detection of outliers is the important field of data mining for which various algorithms have been used [24].In literature, various outlier detection algorithms are used for the detection of the outliers in the given set of data [12].

This paper is organized in following sections after this introduction. Section II represents the work done in the field of outlier detection; it basically represents the literature survey. Section III discuss about categorical classification of existing algorithms that are used for the detection of outliers from the datasets. Section IV represents the comparison of the existing outlier detection algorithms in terms of their efficiency, computational cost, scalability etc. The conclusion by reviewing the various outlier detection techniques and the advancements need to be described in this section.

2. LITERATURE SURVEY
As the tremendous data is increasing day by days so the proper detection of the outliers is becoming challenge. Various renowned researchers [3, 6, 11, 18 & 37] are already working in this area. The research work done by researchers is discussed in this section.

Aggarwal [31] expressed the use of depth based outlier detection method. The proposal was that the points lie at the corners of the convex hull in the outer boundaries of the data. These points in the outer boundaries are identified as likely to be outliers.

Struyf et al., [3] discussed the main drawback of convex hull-based approach in the depth based outlier detection is the

large execution time. It makes impractical for processing large data sets with depth based algorithm.

Rajaraman et al., [2] proposed an excellent recap of the clustering strategies as well as a discussion about the curse of dimensionality on high dimensional sets. It was very relevant because one of the main drawbacks of outlier detection is the degradation of the performance when incrementing the data size.

Liu et al., [6] combined both local and global outlier detection and proposed a method which can clearly handle data having imperfect labels and enhanced the performance of outlier detection. Various experiments have been done on real life data sets which concluded that these proposed approaches can attain better tradeoff between false alarm rate and detection rate as compared to the traditional techniques.

Sreevidya et al., [34] mentioned that assumption based methods are fairly fit if the prior assumption made about the data is right, so in this case prior knowledge of data is required and if it is not known then better to use combined approach for outlier detection. The researchers also observed that efficiency of outlier detection is directly proportional to the type of data and data distribution. They also concluded that individual methods are not much effective over data streams.

Montes [26] have presented a novel approach of Depth based to detect outliers based on squared-well. This approach has been designed to be able to deal with large data sets, at the same time that it keeps a reasonable execution time. It is concluded that this algorithm of the detection of the outlier is much effective than the previous one of depth based.

Barmad et al., [4] presented the various transaction specific outlier detection methods that is used for the detecting of outlier from the database. The identification of the outlier pattern is one of the rising problems. Credit card fraud detection and network intrusion are example of the problem. Previously the outlier detection was done from the numerical data set but problem in this outlier detection was that it was not applicable for the live transaction data base. Outlier detection methods are classified into transaction specific and non transaction specific.

Gupta et al., [23] focused on outlier detection techniques for temporal data, considering various approaches for different applications. Various methods for the temporal outlier detection from the computational perspective were discussed. The survey of the outlier detection in large sets of temporal data was presented.

Christy et al., [1] suggested a technique in which two algorithms are used for the detection of the outliers. These two algorithms are Distance based outlier detection and cluster based outlier detection algorithms which are used for the detection and removal of the outliers. This technique of Outlier detection can be used for various domains like big data, high dimensional data etc. According to researchers, this technique is better for the dimension reduction.

Zhangn et al., [21] proposed the approach for increasing the accuracy of the traditional technique of the outlier detection. It focused on the detection accuracy in the high dimensional data. First the angles between the all pairs of the two lines are assessed after that the outliers are detected by the use of normalized mahalanobis distance. In this artificial high dimensional dataset is created, by doing this the accuracy of the system is increased from the experiments performed and it is concluded that this algorithm is suitable for the fault detection tasks and having various other merits.

Devi et al., [32] suggested the clustering technique of outlier detection for the high dimensional dataset. This technique helped in finding the entities that are different, unique and unmatched with the data that is present in the input dataset. To increase the dimensionality of the dataset connecting to the nearest neighbor, the concept of hub was developed. Clustering plays an important role in handling the high dimensional data and is one of the techniques of the outlier detection. Author used the clustering techniques such as the K-Mean and Fuzzy C mean for the detection of the outliers. It also results in decrease of the computational time. The anti-hub point is embedded in the clusters that are formed after using this clustering method. It is concluded that the anti-hub that is applied into K–Means is more efficient than the anti-hub applied to Fuzzy C Means. A Comparison between both the algorithms is performed that shows the computational time of Acanthus is less than the FC Anti-hub.

Aggarwal et al., [8] focused on the behavior of projections from the datasets and introduced the new technique for the outlier detection. As outlier detection is used in various applications such as detection of the fraud, robustness analysis etc. in the high dimensional data. For this many algorithms have been developed to provide the best technique of the outlier detection. Detection of outliers has become more complex in case of high dimensional data.

Behera [12] suggested a clustering based technique which used K-Mean clustering algorithm for clustering of the datasets and density based and distance based algorithms for finding out the outliers. He concluded that from the experimental analysis of both lower dimension datasets and higher dimension datasets, as in case of Bupa Dataset, K Mean clustering algorithm can be used for outlier analysis.

3. OUTLIER DETECTION ALGORITHMS

Data mining is a useful technique for learning and extracting useful data from a dataset. Various techniques of data mining are designed that helps to search data from the large data set present in the computer [10].The algorithms that are designed should be highly capable, faster, understandable, robust etc. One of the basic problems of data mining is the outlier detection. An outlier is an observation that is quite dissimilar from the other observations. As various techniques have already been introduced till now, in this section different existing outlier detection techniques have been discussed that are used for detection and removal of outliers.

3.1 Statistical Outlier Detection

This technique of outlier detection frames the model simply with the help of the data points that are available for processing. In this field most of the outlier research has been done, due to which many distributions of the data is known. Most of the statistical model can only handle one attribute and those that can handle multi attributes can handle data efficiently up to the K<4. This is completely subjective to the distribution used.

On the basis of the statistics two methods have been described for the outlier detection [36]:

3.1.1 Distribution Outlier Detection

In this method the standard distribution such as normal and Poisson are chosen as the best data on which the data distribution is dependent [19]. The outlier is detected on the

basis of the probability distribution. Method of detecting outliers based on the general pattern within the data points is considered quite efficient.

3.1.2 Depth Based Outlier Detection

In this method th**e** data is referred as the point in the space and is assigned as depth A. K-d space is used for representing the each data object and accordingly the depth is assigned. The detection of the outlier is done on the basis of the depth. The data objects that have smaller depth have high probability of being considered as outlier .This depth based approach is considered to be better than the various other approaches as it solves the problem of the distribution fitting. It allows the processing of the multidimensional data objects. This basic relies on the computation of the K-dimensional convex hull. Although, it is conceptually used for the high dimensional data, but practically it is not applicable for the high dimensional data [20]

3.2 Distance Based Outlier Detection

This is one of the algorithms of outlier detection that is dependent on the distance between the points. The neighbors of the point are selected and checked in this method [5]. If the neighboring points are close then it is considered normal, but if the neighboring point is far away then that case is considered as unusual. This technique is quite efficient as there is no need of defining the explicit distribution that defines the peculiarity [22].

3.3 Cluster Based Outlier Detection

This outlier detection technique is quite effective as the data from the datasets is firstly segmented into clusters. In every cluster each data point is authorized as a degree of the membership [13]. The outlier is detected without any interference in the clustering process. Various clustering approaches are used for the outlier detection [11]. Clustering on streaming data is categorized by grid based and k means/k median methods [36]. In Partitioning methods, various centroid based methods, k means, PAM (Partitioning Around Medoids), CLARA (Clustering LARge Applications) and CLARANS (Clustering Large Applications based on RANdomized Search) etc methods are used [37]. One of the clustering is hierarchical clustering. In it, the whole data set is further decomposed into different small datasets. It is further divided into two categories i.e., Agglomerative methods (in which sample units are combined to form single cluster) and divisive methods (in which single parent cluster is further partitioned) [38].

3.4 Density based Outlier Detection

In this method of detecting the outlier each object that is present is accredited a LOF (Local Outlier Factor). The local outlier factor is basically the degree assigned to object. In this technique, it is checked that how the object is isolated from its neighbor on the basis of the local outlier. The object with high local outlier factor is termed as outlier and the objects having low local outlier factor are considered to be normal. The high local outlier filter depicts the high probability of being outlier [21].

3.5 Parametric and Non- Parametric Outlier Detection

Parametric method for the outlier detection is the result of the various modifications that are done in the various algorithms. It can also be result of the optimization algorithm. The parametric algorithms depend on the data complexity not on the data size [12]. Regression method is one of the parametric outlier detection techniques. This methods help in finding the dependency of the one or more variable on the other variable. This technique is accurate if the user is having a prior knowledge about the data. But for some applications this is not suitable. This type of outlier does not make any prediction about the statistical distribution of the data. The most widely used non parametric technique for the outlier detection is the histograms and kernel density function. The histogram technique is simply used for maintaining the profile of the data .This technique is applicable for the single feature data.

3.6 Classification Based Outlier Detection

This is basically the non- parametric and model based approach that is used for extracting the hidden features and has ability of the learning large complex data. Both multi-class and one class outlier detection techniques are available. One class support vector machines are one of the classifications of outliers. Multiclass availability of outliers is very few. Distributed strategies for multi-class classification needs to be developed. Various classification based outlier detection are Neural Networks, Support Vector Machines and Bayesian Network and in rule based classification, Decision trees [8, 15 & 37] and association rules [30]. When the neural network is used the entire data set is traversed various times to get the accurate model. These neural methods are further categorized into two types [28]

3.6.1 Supervised Neural Methods

In this process for learning process of the data, the classification of the data is done.

3.6.2 Unsupervised Neural Methods

In it, the network consists of the nodes that will depict the portion of the data set.

3.7 Sliding Window Based Outlier Detection

In this technique, the concept of sliding window is used for the streaming data. For the detection of the outlier within the streaming data various algorithm have been designed. The window size is to be accurately chosen as required; also the selection of the sliding window is not dependent on the data point. Some outliers were considered as inliers in other window, so this method is not efficient. Main problem is that sometimes outlier point may get classified as inliers [27].

3.8 DSS and LDSS Outlier Detection Algorithms

DSS (Distributed Solving Set) and LDSS (Lazy Distributed Solving Set algorithms) [10] are distributed strategies for outlier detection in large data sets. DSS algorithm had supervisor node, core computation handled simultaneously by other nodes and synchronization of partial result after completion of the job. Run time of the algorithms is small because number of local nodes is large.

In Lazy DSS, the subsets of collection of nearest neighbors of each candidate node are computed by starting the smallest ones and sending them into the local nodes to the supervisor node.

4. COMPARSION OF EXISTING OUTLIER DETECTION ALGORITHMS

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

of data used, size of the dataset etc [33]. Now with the tremendous growth of data, best outlier detection algorithms have to be applied on large data sets [8]. So the basic parameters like efficiency, computational cost, scalability and applicability needs to be studied.

4.1 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.

Table-1 Comparison of Outlier Detection Techniques

Sr. No.	Algorithms	Efficiency	Computational Cost	Scalability	Application	High Dimensional Data
1	Statistical Based Outlier Detection	Low	High	No	Statistical Data	No
2	Depth Based Outlier Detection	Low	High	No	Statistical Data	No
3	Distance Based Outlier Detection	Average	Low	Yes (but not much)	Based on the distance of individual points	Yes
4	Density Based Outlier Detection	High	High	Yes(High)	Local neighborhood of the data points	Yes
5	Clustering Based Outlier Detection	Very High	Low	Yes(High)	Depends upon the clusters of data	Yes
6	Classification Based Outlier Detection	Very High	Very Low	Yes(High)	Normal training data	Yes
7	Sliding Window Based Outlier Detection	Less	High	Yes	Streaming Data	Yes
8	DSS & LDSS Outlier Detection	High	Very Low	Yes	Large datasets	Yes

4.2 Computational Cost

It is directly proportional to the computational complexity of the algorithm. It is the evaluation 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.

4.3 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.

4.4 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.

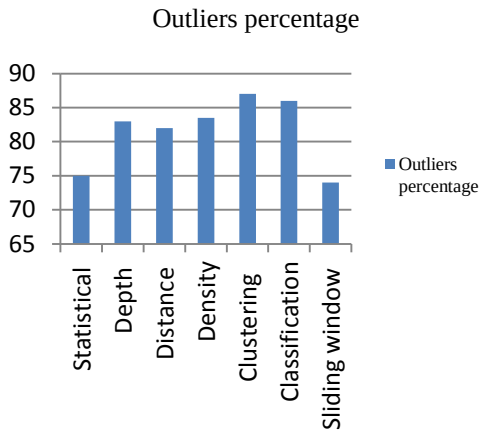

Figure-2 Comparison of different techniques

The above said figure-2 is designed by evaluating the different outlier detection techniques on few huge datasets. The given figure presents their maximum cases of detecting outliers. Clustering and classification algorithms are having a higher percentage of detecting the outliers. It is also seen that there may be variations in the results according to the category of data and how it is being retrieved.

5. CONCLUSION

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. This paper presents 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. It is found that efficiency and computational complexity depends upon the data distribution and type of data. It is also observed that no individual algorithm is much suited for the high dimensional data. There is need of developing some new algorithms or improvement in the existing one is required. In future work, discussed algorithms will be explored using different parameters which are not included in this paper.

6. 6. REFERENCES

[1] A Christy, G. Meera Gandhi, "Cluster Based Outlier Detection Algorithm for Healthcare Data", Elsevier, Volume 50, 2015, pp 209-215.

[2] A. Rajaraman, J.D. Ullman, "Mining of massive datasets", Cambridge UniversityPress, Cambridge, 2012.

[3] A. Struyf, P J Rousseeuw, "High-dimensional computation of the deepest location. Computational Statistics and Data Analysis", American mathematical soc, Volume 34, 2000, pp 415–426.

[4] Anjali Barmad, Madhu M.Nashipudinath, "An Efficient Strategy to Detect Outlier Transactions", International Journal of Soft Computing and Engineering, Volume 3, Issue 6, January 2014, pp 174-178.

[5] B. Wang, Gang Xiao, Hao Yu, Xiaochun Yang, "Distance-Based Outlier Detection on Uncertain Data", IEEE International conference on Computer and Information Technology, Volume 1,October 2009, pp 293 – 298.

[6] Bo Liu, Yanshan Xiao, P.S. Yu, Zhifeng Hao, Longbing Cao, "An Efficient Approach for Outlier Detection with Imperfect Data Labels", IEEE Transactions on Knowledge and Data Engg, 2014, pp 1602 – 1616.

[7] Charu C. Aggarwal, Philip S, "Outlier Detection for High Dimensional Data", In proceedings of ACM SIGMOD International conference on Management of data, Volume 30, Issue 2, June 2001, pp 37-46.

[8] D. Xiang, W. Lee, "Information-theoretic measures for anomaly detection", In proceedings of IEEE Symposium on Security and Privacy, May 2001,pp 130-143.

[9] Dr. T. Christopher, 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, March 2015, pp 194-197

[10] Dragoljub Pokrajac, Aleksandar Lazarevic, Longin Jan Latecki, "IEEE Symposium on Computational Intelligence and Data Mining (CIDM)", April 2007.

[11] F. Anguiulli, F. Fassetti, "Detecting Distance-Based Outliers in Streams of Data", In Proceedings of the 16th ACM Conference on information and knowledge management (CIKM), 2007, pp 811 – 820.

[12] Fabrizio Angiulli, Stefano Basta, Stefano Lodi, and Claudio Sartori, "Distributed Strategies for Mining Outliers in Large Data Sets", IEEE Transactions On Knowledge And Data Engineering, Volume 25, No.7, July 2013, pp 1520-1532.

[13] H.S.Behera, "A New Hybridized K-Means Clustering Based Outlier Detection Technique For Effective Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012, pp 287-292.

[14] HaNguyen Thi Thu,Quynh Nguyen Huu, Tu Nguyen Thi Hgoc, "A Supervised Learning Method Combine With Dimensionality Reduction in Vietnamese Text Summarization", IEEE conference on Computing, Communications and IT Applications (ComComAp), April 2013, pp 69-73.

[15] Helmer G, Wong J, Honavar V, Miller L, "Intelligent agents for intrusion detection", In Proceedings of IEEE Information Technology Conference, July 2013, pp 121-124.

[16] Jayanta K. Dutta, Bonny Banerjee, Chandan K. Reddy, "RODS: Rarity based Outlier Detection in a Sparse Coding Framework", IEEE Transactions on Knowledge and Data Engineering, Volume 28, Issue 2, September 2015,pp 483-495.

[17] Ji Zhang, "Advancements of Outlier Detection: A Survey", ICST Transactions on Scalable Information Systems, Volume 13, Issue 01-03, March 2013, pp 1-16.

[18] Jiadong Ren,"Efficient Outlier Detection Algorithm for Heterogeneous Data Streams", Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009.

[19] K. Bhaduri, "Algorithms for speeding up distance-based outlier detection", In Proceedings of ACM SIGKDD International Conference KDD, New York, USA, 2011, pp 859–867.

[20] Karanjit Singh, "Outlier Detection: Applications and Techniques", IJCSI International Journal of Computer Science Issues, Volume 9, Issue 1, No 3, January 2012, Pp 307-323.

[21] Liangwei Zhangn, Jing Lin, Ramin Karim, "An anglebased subspace anomaly detection approach to highdimensional data: With an application to industrial fault detection" , Elsevier, 2015, Pp 482-497

[22] Mahito Sugiyama, "Rapid Distance-Based Outlier Detection via Sampling", Advances in Neural Information Processing Systems 26 (NIPS 2013), 2013.

[23] Manish Gupta, Jing Gao "Outlier Detection for Temporal Data: A Survey", IEEE transactions on knowledge and data engineering, Vol. 25, No. 1, January 2014, pp 1-20

[24] ManzoorElahi, "DB-Outlier Detection Algorithm using Divide and Conquer approach over Dynamic", International Conference on Computer Science and Software Engineering DataStream", 2008.

[25] MaysoonAbulkhair, "Intelligent Integration of Discharge Summary: a Formative Model", 4th International Conference on Intelligent Systems, Modelling and Simulation, 2013.

[26] Miguel Cardenas Montes, "Depth based outlier detection algorithm", Springer, 2014, pp 122-132.

[27] Ms. S. D. Pachgade, "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, pp 12-16.

[28] P.Chandore, P.Chatur, "Outlier Detection Techniques over Streaming Data in Data Mining: A Research Perspective", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-1, March 2013.

[29] Parneeta Dhaliwal, "A Cluster-based Approach for Outlier Detection in Dynamic Data Streams (KORM: k-median OutlieR Miner)", Journal of computing, volume 2, Issue 2, february 2010, pp 74-80.

[30] Prakash Chandore, "Outlier Detection Techniques over Streaming Data in Data Mining: A Research Perspective", International Journal of Recent Technology and Engineering, Volume-2, Issue-1, March 2013, pp 157- 162.

[31] R Aggarwal, R Srikant, "Mining sequential patterns", In Proceedings of 11th International Conference on Data Engineering IEEE Computer Society, Washington, DC, USA, 1995, pp 3-14.

[32] R.Lakshmi Devi, Dr. R.Amalraj "An Efficient Unsupervised Cluster based Hubness Technique for Outlier Detection in High dimensional data", International Journal of Innovative Research in Advanced Engineering , Volume 2 , Issue 10 , October 2015, pp 63-70.

[33] Rajendra Pamula, Jatindra Kumar Deka, Sukumar Nandi, "An Outlier Detection Method based on Clustering", Second International Conference on Emerging Applications of information Technology, 2011, pp 253-256.

[34] S S Sreevidya, "A Survey on Outlier Detection Methods", International Journal of Computer Science and Information Technologies, Volume 5 (6), 2014, pp 8153-8156.

[35] Shuwu, "Information-Theoretic Outlier Detection for Large-Scale Categorical Data", IEEE transactions on knowledge and data engineering, Volume 25, No. 3, March 2013.

[36] Vijay Kumar, "Outlier Detection: A Clustering-Based Approach", International Journal of Science and Modern Engineering (IJISME), Volume-1, Issue-7, June 2013, pp 16-19.

[37] W. Fan, M. Miller, S. Stolfo, W. Lee, P. Chan "Using artificial anomalies to detect unknown and known network intrusions", In Proceedings of IEEE International Conference on Data Mining, IEEE Computer Society, Volume 6, Issue 5, April 2004, pp 507-527.

[38] Y-Shi, "COID: A cluster- outlier iterative detection approach to multi-dimensional data analysis", Knowledge Information System Volume 288, No 3, 2011, pp 709 – 733.