ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

An Adaptive Parallel Algorithm for Outlier Detection Using Ranking Strategy

Jitendra R. Chandvaniya,
M.Tech (CE) Research Scholar, R K University, Rajkot, India

Manuscript Info

Manuscript History:

Received: 23 April 2014
Final Accepted: 25 May 2014
Published Online: June 2014

Key words
Data-mining, Outliers, Distance-Based Outlier Detection, Ranking

Corresponding Author

Jitendra R. Chandvaniya,

Abstract

Outlier Detection is very much popular in Data Mining field and it is an active research area due to its various applications like fraud detection, network sensor, email spam, stock market analysis, and intrusion detection and also in data cleaning. Most widely use of outlier detection in medical diagnostics to detect irregular patterns in patient medical records which could be symptoms of a new disease So, it is very important to detect outlier in large data set very efficiently, here we have discussed various methods for outlier detection like a very simple and easy method is distance-based outlier detection in this method outlier can be detected based on its distance to predefined points in a given data set, find out the nearest neighbor and based on it detect points as outliers, but it is very challenging task to develop such method which is efficiently detect outliers and also can be applicable to large data set effectively. Ranking based outlier detection is the new area in the outlier detection, various ranking strategies apply to detect outliers but no one is much power full which we can apply for large dimensional datasets. So In this paper we have made a survey to implement An Adaptive algorithm using ranking strategies for large dimensional datasets and also to overcome the limitation of existing system.

INTRODUCTION

Data mining is the process of discovering actionable information from large sets of data [1]. Data mining uses mathematical analysis to derive patterns and trends that exist in data. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD) [3]. The term Knowledge Discovery in Database means KDD refers to finding of useful knowledge in database and use this knowledge for high-level applications using data mining methods [2]. Below figure1 describe the KDD process steps [1, 2, and 3].

Make a target data set: selecting a data set, or take a data sample on which discovery is to be performed.

Data cleaning and preprocessing: In this step removal of noise or outliers is done, than collecting necessary information is done, techniques to handle missing data fields and at last Accounting for time sequence and known changes is done.

Transformation: Next is to transform data to particular form to perform data mining.

Data mining: Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.

Interpretation/Evaluation: Interpreting the patterns into knowledge by removing redundant or irrelevant patterns; translating to useful patterns into terms that human understandable.

Data mining is accomplished by building models. A model uses an algorithm to act on a set of data. The notion of automatic discovery refers to the execution of data mining models [1]. Data mining is primarily used today by companies with a strong consumer focus-retail, financial, communication, and marketing-organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables

them to determine the impact on sales, customer satisfaction, and corporate profits [1]. Finally, it enables them to "drill down" into summary information to view detail transactional data. With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

Outlier Detection is very much popular in Data Mining field and it is an active research area due to its various applications like fraud detection, network sensor, email spam, stock market analysis, and intrusion detection and also in data cleaning. The importance of outlier detection is due to the fact that outliers in data trans- late to significant information

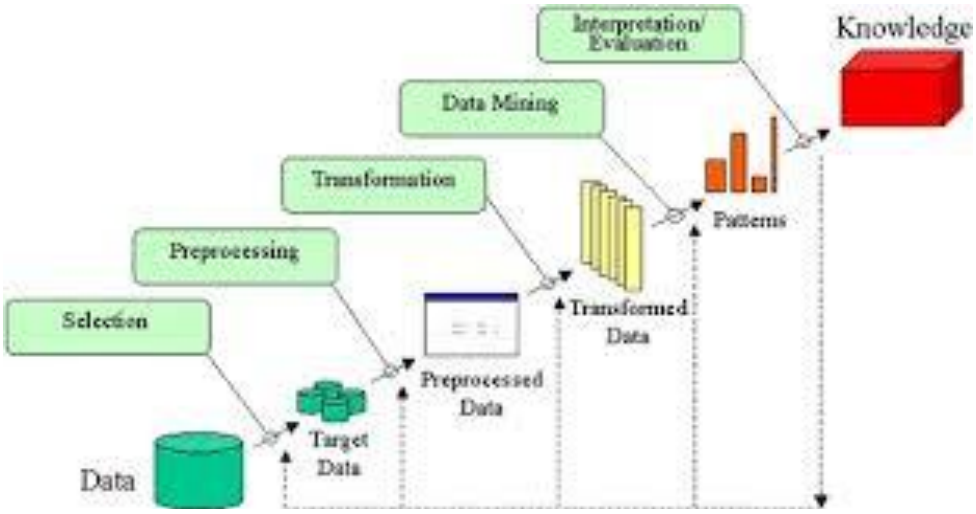

Fig 1: KDD Process Overview [2]

in a wide variety of application domains [4]. For example an unusual traffic design in a computer network might mean that a hacked computer is sending out sensitive data to an unauthorized destination. In public health data, outlier detection techniques are widely used to detect inconsistent patterns in patient medical records which could be symptoms of a new disease. Similarly, outliers in credit card transaction data could indicate credit card theft or misuse. Outliers can also translate to critical entities such as in military surveillance, where the Presence of an unfamiliar region in a satellite image of enemy area could indicate enemy troop movement or anomalous readings from a space craft would signify a fault in some component of the craft [5]. So, it is very important to detect outlier in large data set very efficiently, there are various methods available for outlier detection like a very simple and easy method is distance-based outlier detection in this method outlier can be detected based on its distance to predefined points in a given data set, find out the nearest neighbor and based on it detect points as outliers but it is very challenging task to develop such method which is efficiently detect outliers within less time complexity and which can be apply to large data set effectively. Outliers can be defined as a pattern or characteristic which does not conform to normal behavior of the data set. Outliers have different characteristic then other data in the given data set [5].

Here figure 2 gives basic idea about the outliers, here in the above figure data distribution is given in this one of them have different data value then other in the distribution so such data values in the data set can be defined as Outliers. Outliers exist in almost every real data set. Some of the prominent causes for outliers are listed below [4].

- **Malicious activity**: such as insurance or credit card or telecom fraud, a cyber-intrusion, a terrorist activity

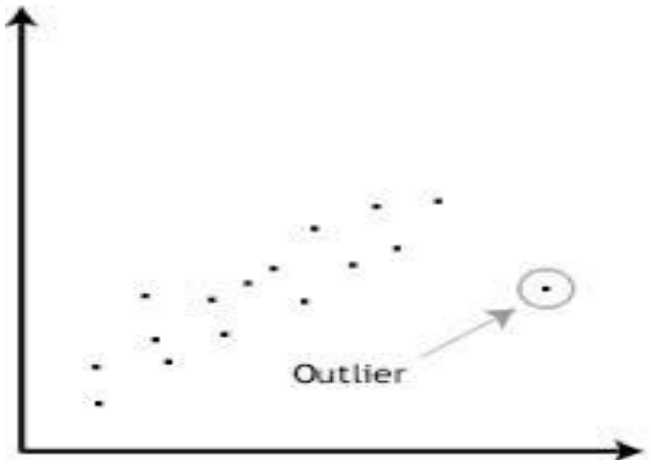

Fig 2: An Example of Outlier [7]

- **Instrumentation error**: such as defects in components of machines or wear and tear
- **Change in the environment**: such as a climate change, a new buying pattern among consumers, mutation in genes
- **Human error**: such as an automobile accident or a data reporting error

Now, what is Outlier Analysis: Outlier analysis can be defined as: Given a set of data objects or data points N and the number of outliers K, find top K outlier points which are considerably dissimilar from the remaining data. The outlier analysis involves defining what data can be considered as inconsistent in a given data set and then to mine the outliers so defined by using efficient techniques [6].There are various techniques available for outlier detection we in this paper we will consider Distance-based outlier detection technique. In this technique based on the distance outliers can be detected, various techniques are used to calculate this distance. In this we need to set some threshold value and based on this threshold value the data in the data set have more than this threshold value can be defined as an outliers. Existing system is useful with low dimensional datasets and also give more false detection in the result and the ranking strategies used in this system does not work with large dimensional datasets Using our simulation codes which is developed in R-Language we can get more accurate plots and graphs which is more reliable to understand the performance of the developed system.

I. RELATED WORK

There are various work done in outlier detection using various techniques in this section we will see one by one techniques, First algorithm based on distance proposed by Edwin M. Knorr and Raymond T. Ng Algorithms for Mining Distance-Based Outliers in Large Datasets [9] In this author determine two algorithms, first one is a nested loop algorithm that runs in O(dN*N)time, on other hand is cell-based algorithm that is linear with respect to N where N is the number of points of the data set, but exponential in d where d is the dimensions of the data set. This method efficiently works if d<=4, while nested loop algorithm is effective work for small data set to be mined. Sridhar Ramaswamy proposed Efficient Algorithms for Mining Outliers from Large Data Sets [10], In this author rank each point on the basis of its distance to *kth* nearest neighbour Outliers detection in this method done using partition-based algorithm, first it partitions the points using clustering algorithm and then prunes those partitions that cannot contain outliers. Wen Jin define Ranking Outliers Using Symmetric Neighbourhood Relationship [11], here author use measure on local outliers based on a symmetric neighbourhood relationship. The proposed measure considers both neighbours and reverse neighbours of an object when estimating its density distribution and then based on it detect top-n outliers. Carlos H. C. Teixeira proposed An Efficient Algorithm for Outlier Detection in High Dimensional Real Databases [12] main aim of this algorithm is a fast strategy to estimate the unusualness of a record within the database and use a rank-ordered approach to evaluate records. Algorithm partitions the database and ranks the objects that are candidates to be an outlier, it reduce the number of comparisons among objects. Author evaluates different ranking heuristics in a wide-ranging set of real and synthetic databases.. Nguyen Hoang Vu and Vivekanand Gopalkrishnan define Efficient Pruning Schemes for Distance-Based Outlier Detection [13] in the first phase, partition the data into clusters, and make an early estimate on the lower bound of outlier scores. Based on this lower bound, the second phase then processes relevant clusters using the traditional block nested-loop algorithm. Here two efficient pruning rules are utilized to quickly discard more non-outliers and reduce the search space, another Approach defined by rajendra is Distance Based Fast Outlier Detection Method [14], in this local

distance-based outlier factor used to measure the degree to which an object departs from its neighbourhood. Then author use pruning strategy to prune out some of the point using clustering algorithm from the given data set which are probably not the member of outliers.

Srinivasan Parthasarathy define Distance Based Outlier Detection: Consolidation and Renewed Bearing [8] In this author use combination of optimization strategies which can give more efficiency, here author use different pruning strategies and ranking strategies and combining them for outlier detection purpose, in this author conclude that combination of ROCO and ANNI is able to achieve one of the best execution times. But for large number of objects like Uniform30D database which has 30 dimensions this algorithm is not applicable.

Bhaduri define Algorithms for speeding up distance-based outlier detection [15], here author introduce sequential and distributed algorithm. Sequential algorithm (iOrca) and Distributed algorithms Door and iDOor, combination with index scheme with distributed processing algorithm works speedily and also applicable for large datasets.Srinivasan Parthasarathy proposed Locality Sensitive Outlier Detection: A Ranking Driven Approach [16], here author develop a light-weight ranking scheme that is driven by locality sensitive hashing, which reorders the database points according to their likelihood of being an outlier. Here ranking scheme improves the effectiveness of the distance-based outlier detection process by up to 5-fold. Ms. S. D. Pachga defined Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach [18], in this author use combination of cluster-based and distance based outlier detection

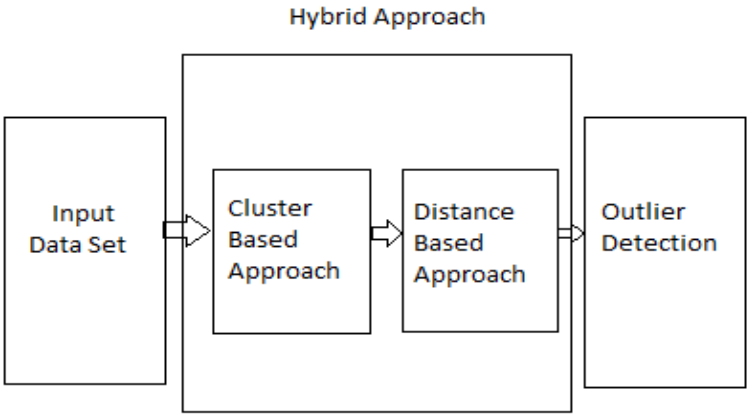

Fig 3: Hybrid Approach model

This approach deals with only numerical data and it cannot deal with more complex datasets.Yanyan Huang define A Hybrid Distance-Based Outlier Detection Approach [17], in this author uses average distance as neighborhood distance, and record the number of data object points within the neighborhood, so that average number of neighbors can be calculated. . Vijay Kumar define algorithm for detection of outlier using cluster-based approach [19] in this approach first they do Partition Around Medoids (PAM) clustering algorithm. After that small clusters are determined and consider as an outlier cluster. Then using absolute distances between the medoid of the current cluster and each one of the points in the same cluster, through this calculation detecting the rest of the outliers (if any).H.Huang defines "Rank-based Outlier Detection" algorithm In this author propose new approach for outlier detection [20], based on this new ranking measure the purpose of this measure is whether a point is "important" for its nearest neighbor. Here author use notation low cumulative rank which says that the point is central. Centrally located point in a cluster has relatively low cumulative sum of ranks because it is among the nearest neighbor of its own nearest neighbors. So, rank measures an object's outlierness. Sum of ranks of an object is naturally meaningful to measure the degree of isolation of an object.

II. RANKING APPROACHES

1. ROCN (Ranking Object Candidates for Neighbors):

This strategy ranks the order in which neighbors of points are processed so this will reduce the current value of $D^k(p)$ faster. It reorders the neighboring clusters so that the search for neighbors proceeds from closest to distant, so we can say that ROCN helps in improving the performance of neighbor while evaluating a given point. Ranking the

search for neighbors in a partition level using estimated distances among them. These estimates could be determined by calculating the distances between the centers (centroids) of the partitions or even between MBR structures
As above we have seen various papers in which they have used ROCN strategy, we find that ROCN is significant and able to improve the execution time but when we apply it to large dimensional dataset like Uniform30D this strategy not gave best execution time [6, 7, 8, 11, and 14].

2. ROCO (Ranking Object Candidate for Outlier):

This strategy used to decide which objects are more likely to be outlier. Aim of this strategy is to focus on the value of D^k min. The strategy estimate the *kth*-NN distance for each Object p. These estimates are then used as a ranking where in objects with greater *kth* NN distances will be considered first as candidate outliers. Ranking objects that are candidates for outliers this strategy is more likely to density-based heuristic, in which the intuitions that have low-density regions (partitions) tend to contain higher-score objects. We define density as |P|/R (P), where |P| is the number of objects in partition P, and R (P) is the MBR diagonal length of P.
As above we have seen various papers in which they have used ROCO strategy, we find that this strategy is very much use full and give much better result compare to ROCN but still it not much effective with large dimensional dataset [11, 14, and 19].

3. OTHER METHODS OF RANK BASED OUTLIER DETECTION

1. LOF(Local outlier Factor) approach

In this approach author proposed that each data point of the given data set should be assigned a degree of outlines and they refer it as the "Local Outlier Factor" (LOF) [20, 22] of the data point and it is calculated as given below.

$$L_k(p) = [\sum o \epsilon Nk(p) \frac{lk(0)}{lk(p)} lk(o)/|Nk(p)|]$$

$L_k(p)$ is calculated for selected values of k in a pre-specified range, max $L_k(p)$ is retained, and a p with large LOF is declared to be outlier [20]

2. COF(Connectivity-based outlier factor) approach

This is modified technique of LOF this says that when a cluster and a neighboring outlier have similar neighborhood densities. COF can be measure as given below formulae.

$$COF_k(p)=[A_{Nk(p)}(P)[\sum_{o\epsilon Nk(p)} A_{Nk(p)}(o)/|N_k(p)|]^{-1}$$

Here larger value of $COF_k(p)$ denotes higher Possibility that p is an outlier [20, 21].

3. INFLOW(INFLuential measure of outlier by symmetric relationship approach)

This method is based on the concept of symmetric neighborhood relationship in this considers neighbors and reverse neighbors of a data point when estimating its density distribution. Using this outlier can be detected as given below method.

$$INFLO_k(p) = 1/den(p) * \sum_{o\epsilon ISk(p)} den(o)/|(IS_k(p))|$$

Where den $(p) = 1/d_k(p)$ [20, 21].

III. SIMULATION ENVIRONMENT

Here for implementation purpose I have used R-Language; R is an integrated suite of software facilities for data manipulation, calculation and graphical display [21]. Among other things it has
1. An effective data handling and storage facility,

2. A suite of operators for calculations on arrays, in particular matrices,
3. A large, coherent, integrated collection of intermediate tools for data analysis,
4. Graphical facilities for data analysis and display either directly at the computer or on hardcopy, and
5. A well developed, simple and effective programming language (called 'S') which includes conditionals, loops, user defined recursive functions and input and output facilities. (Indeed most of the system supplied functions are themselves written in the S language.) The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software [21, 22]. R is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been ex-tended by a large collection of packages. However, most programs written in R are essentially ephemeral, written for a single piece of data analysis [23].

IV. EXPERIMENTAL EVOLUTION

Here for implementation purpose we have used human brain dataset as an input dataset to find outliers from it and rank than according to distance-based. Few packages i have used for this implementation.
DMwR: Data mining with R,this package used for data mining terms and functions [26].
mvoutlier: Multivariate outlier detection based on distance-based methods, this package used for high-dimensional. Outliers detection purpose using distance-based technique and also provide distance so we can easily find outliers based on distance and can rank them [27].
gplots: For representation of result with plotting.
graphics: for graphical purpose like adding color in the graphs or plots to represent result.

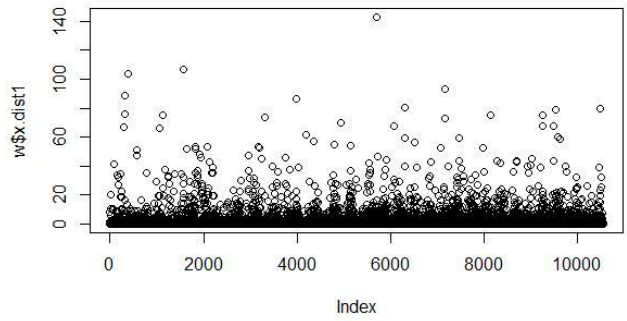

Fig 4: numeric vector with distances for location outlier detection

Above figure 4 shows the plot of numeric vector with distance for location outlier detection, high distance can be the most eligible candidate for outliers.

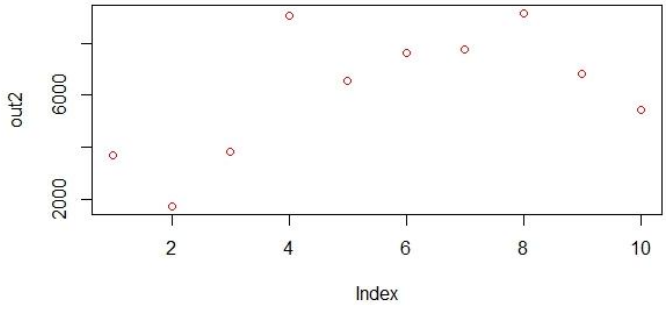

Fig 5: Top 10 outliers in decreasing order

Above figure 5 shows the top 10 outliers and rank of those outliers' shows in decreasing order. So, this solves the problem with existing problem.

V. CONCLUSIONS

After reviewing all this papers which are concerned with my base paper as well as outlier detection area of data mining we can conclude that right now much more research field is covered by previous scholars in outlier detection but ranking-based outlier detection is a new concept in this approach some research work done but that not satisfied the various challenges of outlier detection so, it is needed to develop new approach using ranking strategies to find outliers. Here we have implemented a program using R language's in built functions which can Detect outliers and show some outputs of it. It can also find the distance between the neighbors through which we can give ranking, large distance have more priority and so on.

VI. FUTURE WORK

The system that I have implemented that is able to detect outliers and also find the distance how they are differ from its neighbor and rank them in decreasing order so, the maximum difference object stand first it has higher priority as an outlier. In future we can develop system which can remove the top-n outliers from the dataset that we have detected and use the remaining dataset for useful purpose so that we can apply it for various applications.

VII. REFERENCES

[1] Han J. and Kamber M. (2006),"'Data Mining Concepts and Techniques"', San Fran- cisco, CA, Elsevier Inc.

[2] Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, Advances in Knowledge Discovery and Data Mining,AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34

[3] Introduction to Data Mining and Knowledge Discovery, 3rd Edition ISBN: 1- 892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.Ch.1

[4] VARUN CHANDOLA University of Minnesota: Outlier Detection : A Survey

[5] en-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) "`Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers"', Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2

[6] CHARU C. AGGARWAL IBM T. J. Watson Research Center, Yorktown Heights, NY, USA: OUTLIER ANALYSIS

[7] Allen B. Downey "Think Stats" ,Publisher:O'Reilly Media, Inc,Pub. Date: July 15, 2011,Print ISBN-13: 978-1-4493-0711-0.

[8] Gustavo H. Orair Carlos H. C. Teixeira Wagner Meira Jr., Ye Wang SrinivasanParthasarathy "DistanceBased Outlier Detection: Consolidation and Renewed Bearing" Proceedings of the VLDB Endowment, Vol. 3, No. 2 Copyright 2010 VLDB Endowment 21508097/10/09

[9] E. Knorr and R. Ng. "`Algorithms for Mining Distance-Based Outliers in Large Datasets"'. In Proceedings of VLDB'98, pages 392-403, 1998.

[10] S. Ramaswamy, R. Rastogi, and K. Shim. "`Efficient Algorithms for Mining Outliers from Large Data Sets"'. In Proceedings of SIGMOD'00, pages 427-438, 2000

[11] Wen Jin1, Anthony K. H. Tung2, Jiawei Han3, and Wei Wang4," Ranking Outliers Using Symmetric Neighborhood Relationship", 2006

[12] Carlos H. C. Teixeira, Gustavo H. Orair, Wagner MeiraJr, SrinivasanParthasarathy "An Efficient Algorithm for Outlier Detection in High Dimensional Real Databases",2008

[13] Nguyen Hoang Vu and VivekanandGopalkrishnan," Efficient Pruning Schemes for Distance-Based Outlier Detection", W. Buntine et al. (Eds.): ECML PKDD 2009, Part II, LNAI 5782, pp. 160-175, 2009

[14] RajendraPamula,Jatin, "Distance Based Fast Outlier Detection Method",987-1-4244-9074-5/10,26.00 2010 IEEE, India Conference (IN- DICON), 2010 Annual IEEE.

[15] KanishkaBhaduri,Bryan L. Matthews," Algorithms for speeding up distance-based outlier detection", SIGKDD(special interest group on Knowledge Discovery and Data mining),2011

[16] YeWang, SrinivasanParthasarathy, ShirishTatikonda," Locality Sensitive Outlier Detection: A Ranking Driven Approach" Computer Science and Engineering Department, The Ohio State University, OH, USA,2011

[17] Yanyan Huang, Zhongnan Zhang*, Minghong Liao, Yize Tan, ShaobinZhou," A Hybrid Distance-Based Outlier Detection Approach, 2012 Inter- national Conference on System and Informatics(ICSAI 2012),987-1-4673-0199-2/12,31.00 2012 IEEE

[18] 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,2012

[19] Vijay Kumar, Sunil Kumar, Ajay Kumar Singh," Outlier Detection: A Clustering-Based Approach", International Journal of Science and Modern Engineering (IJISME), ISSN: 2319-6386, Volume-1, Issue-7, June 2013

[20] H. Huang, K. Mehrotra, C. K. Mohan," Rank-Based Outlier Detection", Electrical Engineering and Computer Science Technical Reports. Paper 47,2011

[21] An Introduction to R, Notes on R: A Programming Environment for Data Analysis and Graphics,Version 3.0.1 (2013-05-16)

[22] R - Statistical and Graphical Software Note, School of Mathematics, Statistics and Computer Science University of New England, R programming guide Printed at the University of New England, December 12, 2005

[23] R and Data Mining: Examples and Case Studies 1 Yanchang Zhao, yanchang@rdatamining.com http://www.RDataMining.com April 26, 2013

[24] Package 'outliers',August 29, 2013

[25] Luis Torgo LIACC-FEP "`Data Mining with R"'learning by case studies,University of Porto R. Campo Alegre, ,May 22, 2003

[26] Package 'DMwR',August 29, 2013

[27] Package 'mvoutlier',February 25, 2014