Different Outlier Detection Algorithms in Data Mining: A Review

Amandeep Kaur[1] Ms. Kamaljit Kaur[2]
[1]M.Tech Student [2]Assistant Professor
[1,2]Department of CSE
[1,2]Sri Guru Granth Sahib World University Fatehgarh Sahib, Punjab , India

Abstract— Outlier is defined as an observation that deviates too much from other observations. The identification of outliers can lead to the discovery of useful and meaningful knowledge. Outlier detection has been extensively studied in the past decades. However, most existing research focuses on the algorithm based on special background, compared with outlier detection approach is still rare. Most sophisticated methods in data mining address this problem to some extent, but not fully, and can be improved by addressing the problem more directly. The identification of outliers can lead to the discovery of unexpected knowledge in areas such as credit card fraud detection, calling card fraud detection, discovering criminal behaviors, discovering computer intrusion, etc. This paper mainly discusses and compares approach of different outlier detection from data mining perspective, which can be grouped into statistical-based approach, distance-based approach, density-based approach, Information theoretic-based approach.

Keywords: Outlier detection, Statistical-based approach, Distance-based approach, Density-based approach, Information theoretic-based approach.

I. INTRODUCTION

Data mining is a process of extracting valid, previously unknown, and ultimately comprehensible information from large datasets and using it for organizational decision making [1]. However, there a lot of problems exist in mining data in large datasets such as data redundancy, the value of attributes is not specific, data is not complete and outlier [2].

Outlier is defined as an observation that deviates too much from other observations that it arouses suspicions that it was generated by a different mechanism from other observations [3]. The identification of outliers can lead to the discovery of useful and meaningful knowledge and has a number of practical applications in areas such as transportation, public safety, public health and location based services. Recently, a few studies have been conducted on outlier detection for large dataset [4]. However, most existing research focuses on the algorithm based on special background, compared with outlier detection approach is still rare. This paper mainly discusses about outlier detection approaches from data mining perspective. The inherent idea is to research and compare achieving mechanism of those approaches to determine which approach is better based on special dataset and different background.

II. OVERVIEW OF OUTLIER DETECTION

A. OUTLIER

An outlier is an observation point that is distant from other observations. An outlier is due to variability in the measurement or it may indicate experimental error the latter are sometimes excluded from the data set. Outliers being the most extreme observations may include the sample maximum or sample minimum or both depending on whether they are extremely high or low.The sample maximum and minimum are not always outliers because they may not be unusually far from other observations[5].

B. TYPES OF OUTLIERS

In general, outliers can be classified into three categories, namely global outliers, contextual outliers and collective outliers [6], as shown in figure1.

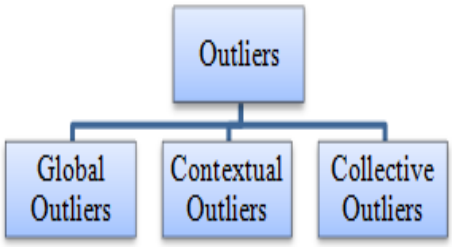

Fig. 1: Types of Outliers

1) Global outlier

In a given data set,a data object is a global outlier if it deviates significantly from the rest of the data set.Global outliers are sometimes called point anomalies and are the simplest type of outliers .Most outlier detection methods are aimed at finding global outlier .For example, Intrusion detection in computer networks [6].

2) Contextual Outliers

In a given data set , a data object is a contextual outlier if it deviates significantly with respect to a specific context of the object. Contextual outliers are also known as conditional outliers because they are conditional on the selected context .Therefore in this kind of outlier ,the context has to be specified as part of the problem definition[6].

3) Collective Outliers

A subset of data objects collectively deviate significantly from the whole data set, even if the individual data objects may not be outliers. Detection of collective outliers, consider not only behavior of individual objects, but also that of groups of objects**.** Need to have the background knowledge on the relationship among data objects, such as a distance or similarity measure on objects [6].

C. OUTLIER DETECTION

Outlier detection encompasses aspects of a broad spectrum of techniques. Many techniques employed for detecting outliers are fundamentally identical but with different names chosen by the authors. For eg, authors describe their various approaches as outlier detection, novelty detection, anomaly detection, noise detection, deviation detection or exception mining. Outlier detection is a critical task in many safety critical environments as the outlier indicates abnormal running conditions from which significant performance degradation may well result, such as an aircraft engine rotation defect or a flow problem in a pipeline[7].

III. OUTLIER DETECTION APPROACH

Outlier detection has been extensively studied in the past decades and numerous approaches have been developed. These approaches can be mainly classified as: Statistical-based approach, Distance-based approach, Density-based approach, Information theoretic-based approach ,as illustrated in Figure 2.

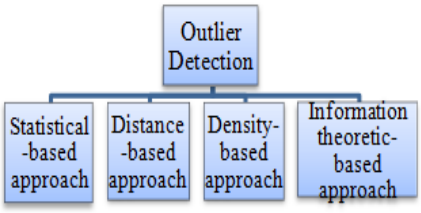

Fig. 2: Outlier Detection Approach

A. Statistical-based approach

Statistical approaches were the earliest algorithms used for outlier detection, which assumes a distribution or probability model for the given data set and then identifies outliers with respect to the model using a discordancy test. In fact, many of the techniques described in both Barnett and Lewis [8] and Rousseeuw and Leroy [9] are single dimensional. However, with the dimensions increasing, it becomes more difficult and inaccurate to make a model for dataset. Statistical approach for outlier detection can be divided into two major categories : parametric methods and non-parametric methods[6].

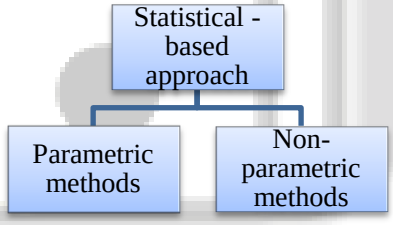

Fig. 3: Methods of Statistical -based approach

1) Parametric methods

Parametric methods assumes that the normal data is generated by a parametric distribution with parameter θ .The probability density function of the parametric distribution $f(x, \theta)$ gives the probability that object x is generated by the distribution. The smaller this value, the more likely x is an outlier [6].

2) Non-parametric methods

Non-parametric method not assume an a-priori statistical model and determine the model from the input data and not completely parameter free but consider the number and nature of the parameters are flexible and not fixed in advance. Examples: histogram and kernel density estimation [6].

3) Control Chart Technique (CCT)

We study control chart technique for outlier data detection. Usually, CCT is used to determine whether your process is operating in statistical control. The purpose of a control chart is to detect any unwanted changes in the process.
These changes will be signaled by abnormal (outlier) points on the graph [25]. Basically, control chart consists of three basic components:

- a centre line, usually the mathematical average of all the samples plotted.
- upper and lower control limits that define the constraints of common cause variations.
- performance data plotted over time.

Firstly, calculate the average for data points to get a centerline of a control chart. Secondly, calculate the upper control (UCL) and lower control limit (LCL). Finally, data are plotted on the chart and data that are out from UCL and LCL and are detected as outlier data.

4) Linear Regression Technique (LRT)

There have been many statistical concepts that are basis for data mining techniques such as point estimation, Bayes theorem and regression. Nevertheless, for this outlier detection analysis, LRT is being used because it is appropriate to evaluate the strength of a relationship between two variables. In general, regression is the problem of estimating a conditional expected value. While as "linear" refers to the assumption of a linear relationship between y (response variable) and x (predictor variable). Thus, in statistics, linear regression is a method of estimating that linear relationship between the input data and the output data [26].

5) Limitation of Statistical -based approach

Statistical-based depend on the parameters of the distribution (Mean and Variance) which are unknown[10].

B. Distance-based Approach

The notion of outliers in DB (p, D) - outlier: An object O in a dataset T is a DB(p, D)-outlier if at least fraction p of the objects in T lies greater than distance D from O. Distance – Based outlier detection using parameters p and D. It is suitable for situations where the observed distribution dose not fit any standard distribution and the very important about Distance-Based outlier, is that it is well defined for *k*-dimensional datasets for any value of *k*. There are two simple algorithms, both having a complexity of O (k N ^{2}), k is dimensionality and N is the number of objects in the dataset. These algorithms readily support datasets with many more than two attributes, and they also present optimized cell-based algorithm that has a complexity that a linear w.r.t. N, but exponential w.r.t *k*, as well as, for datasets are mainly disk-resident, they have another version of the cell-based algorithm [10].

DB(p,D)-outlier detected using parameters p and D.The user has to choose suitable values for p and D to defined the strength of the outliers requested which may involve trail and error and numerous iterations. In this case is quite difficult to choose suitable values for p and D so it will be costly and it dose not provide a ranking for the outliers. For instance a point with very few neighboring points within a distance D can be regarded in some sense as being a stronger outlier than a point with more neighboring within a distance D.Cell-based algorithm whose complexity is linear in the size of the database dose not scale for higher number of dimensions (e.g.,5)[10]. S.Ramaswamy [11] presents a new definition for outliers and developed algorithms for mining outliers which the user does not need to specify the distance parameter D. Instead, it is based on the distance of the *k* nearest neighbor of a point(special case of DB(p,D)-outlier).But ,they have the same weakness

which is they are not powerful enough to cope with certain process with different densities in data clusters[12].

1) Manhattan Distance Technique (MDT)

Commonly, the distances can be based on a single dimension or multiple dimensions. It is up to the researcher to select the right method for his/her specific application. For this outlier detection analysis MDT is used because the data are single dimension. The general formula for MDT is,

$$d(t_i, t_j) = \sum_{h=1}^{k} |(t_{ih} - t_{jh})| \quad (27)$$

Where: $ti = <ti1,...,tik>$ and $tj = <tj1,...,tjk>$ are tuples in a database[27].

2) Limitation of Distance-based Approach

Distance-base, DB (p, D) depends on parameters p (which always close to 1), and the value of D the user has to choose (try by error)[10].

C. Density-based Approach

It assigns to each object a degree to be an outlier. This degree is called the local outlier factor (LOF) of an object. It is local in that, the degree depends on how isolated the object is with respect to the surrounding neighborhood. In LOF algorithm, outliers are data objects with high LOF values whereas data objects with low LOF values are likely to be normal with respect to their neighborhood. High LOF is an indication of low-density neighborhood and hence high potential of being outlier[10].

Papadimitriou et al. [13] present LOCI (Local Correlation Integral) which uses statistical values based on the data itself to tackle the issue of choosing values for MinPts.

1) Advantage of Density-based Approach

Density-based techniques have the advantage that they can detect outliers that would be missed by techniques with a single, global criterion. However, data is usually sparse in high-dimensional spaces rendering density-based methods problematic [14].

2) Limitation of Density-based Approach

Density-base local outlier depends on the value of MinPts, if the value of MinPts is large; LOF has to be computed for every object before the few outliers are detected. This is not a desirable exercise since outliers constitute only small fraction of the entire dataset [10].

IV. INFORMATION THEORETIC-BASED APPROACH

Several information-theoretic methods have been proposed in the literature. For anomaly detection in audit data sets, Lee and Xiang [15] present a series of information-theoretic measures, i.e., entropy, conditional entropy, relative conditional entropy, and information gain, to identify outliers in the univariate audit data set, where the attribute relationship does not need to be considered. The work of He et al. [16] employs entropy to measure the disorder of a data set with the outliers removed. In these methods, heuristic local search is used to minimize the objective function. The methods proposed in [8] and [17] set a threshold of mutual information and obtain a set of dependent attribute pairs. Based on this set, an outlier factor for each individual object is defined. In general, information-theoretic methods focus either on a single entropy-like measurement or on mutual information, and require expensive estimation of the joint probability distribution when the data set is shrunk following elimination of certain outliers.

The two algorithms for outlier detection of Information-Theory- Based approach . One is named ITB-SS for Information-Theory- Based Step-by-Step (or SS for short), the other one is named ITB-SP for Information-Theory-Based Single-Pass (or SP for short). Both algorithms detect outliers one by one [18].

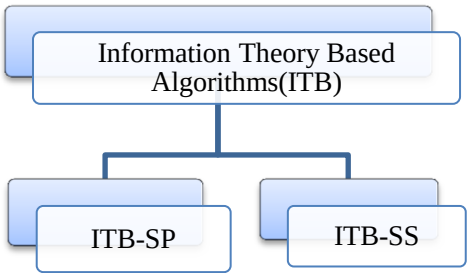

Fig. 4: Information Theory Based Algorithms

In both algorithms, search is conducted only within the anomaly candidate set AS.

A. Information-Theory-Based Single-Pass Algorithm

In Single Pass, the outlier factors are computed only once, and the o objects with the largest OF(ox) values are identified as outliers. In both algorithms, search is conducted only within the anomaly candidate set AS (anomaly candidate set), although this does not make any difference for the algorithm ITB-SP since the initialization of AS requires computation of the outlier factors of all the objects. In ITB-SP, the attribute weights $w_x(y_i)(1 \le i \le m)$, the outlier Factor $OF(x_i)$ of all the objects, initialization of AS and the heapsort search to find the top-o outlier candidates are computed . The time complexity of ITB-SP O(nm). The upper bound on outliers (UO) is to estimate an upper limit on the number of outliers in a data set.

Algorithm 1. ITB-SP single pass
1: Input: data set X and number of outliers requested o
2: Output: outlier set OS
3: Compute $W_x(y_i)$ for$(1 \le i \le m)$
4: Set OS =0
5: for i= 1 to n do
6: Compute $OF(x_i)$ and obtain AS
7: end for
8: if o > UO then
9: o= UO
10: else
11: Build OS by searching for the o objects with greatest $OF(x_i)$ in AS using heapsort
12. endif.

B. Information-Theory-Based Step-by-Step Algorithm

At each step of Step-by-Step, the object with the largest OF(xo) is identified as an outlier and is removed from the data set. Following this removal, the outlier factor OF(x) is updated for all the remaining objects. The process repeats until o objects have been removed.

Algorithm 2. ITB-SS Step-by-Step
1: Input: data set X and number of outliers requested o
2: Output: outlier set OS
3: Set OS =0

4: Compute $W_x(y_i)$ for$(1 \le i \le m)$
5: for i=1 to n do
6: Compute $OF(x_i)$ and obtain AS
7: end for
8: if o > UO then
9: o = UO
10: else
11: for i = 1 to o do
12: Search for the object with greatest OF(xo) from AS
13: Add x0 to OS and remove it from AS
14: Update all the OF(x) of AS
15: end for
16: end if

1) Algorithm

In this algorithm, Considering that o(UO) is usually larger than n, it is possible to say that the final complexity of ITB-SS is O(om(UO)). Compared with ITB-SP, the time complexity of the ITB-SS method is a little higher.

2) Limitation of Information-Theory-Based Algorithm

This technique is works on only unsupervised algorithm. There are required knowledge about statistics [18].

V. IMPLEMENTION

In this section, firstly, we compared the efficiency of the linear regression and control chart techniques (statistical approach). The implementation of both algorithms is using Matlab and Microsoft Access as its database. Through the performance evaluation, we are going to show that the control chart technique is better than linear regression due to the number of outlier data detection is smaller than linear regression technique. This outlier analysis is based on air pollution data. The example of air pollution data is shows in Table 1:

Date	CO	O_3	PM_{10}	NO_2	SO_2
1/8/02	2.26	0.010	74	0.005	0.041
2/8/02	2.46	0.120	68	0.004	0.037
…..	…..	…..	…..	…..	…..
30/8/02	2.05	0.012	60	0.006	0.029

Table. 1: Air Pollution Data[30]

Based on both techniques, outlier data was determined if the data was out of the control limits or boundaries. In control chart technique, UCL and LCL were determined. While as, upper and lower boundaries in linear regression techniques are based on 95 percent computation from liner regression equation that has been identified.

Data	Outlier data for CCT	Outlier data for LRT
CO	16	25
O_3	18	30
PM_{10}	20	25
SO_2	21	29
NO_2	16	21

Table. 2: Results for CCT and LRT

As illustrated in Table 2, outlier data that have been detected by control chart were lower than linear regression technique. This implies that, the lower the number of outlier data detected, the better the technique is. This is due to data plotted on control chart technique are more converged on the data average line. Thus, there are more useful data that could be used for analysis and further could acquire an accurate result.

Secondly, we analysis the MDT (distance-based approach).The implementation of this algorithm also using Matlab and Microsoft Access as its database. In Manhattan distance technique, the threshold values (tv) have to be assigned. Besides that, outlier data also depends on the threshold distance values (d_3). The d3 have to be smaller than maximum distance values (d_2) that exist between each of the data. This is to ensure that d_3 did not out of range and the comparison process could be done. We can get the parameter value (p) by comparing d3 and the distances of each data (d_1). Further, we compare t with p to gain outlier data , we obtained d_2, d_3, tv and the number of outlier as in Table 3.

Data	Max. distance value (d_2)	Threshold distance value (d_3)	Threshold value (*tv*)	Number of Outlier
CO	1.82	1.0	2	15
			4	13
			6	9
O_3	0.08	0.01	2	27
			5	17
			7	11
PM_{10}	81	50	2	7
			3	5
			4	2
SO_2	0.07	0.003	4	21
			5	12
			6	11
NO_2	0.028	0.010	6	12
			7	7
			8	5

Table. 3: Results for MDT

Table 3 show that when the threshold values increases, the number of outlier data detected decreased. This implies that, numbers of outliers are inversed with threshold value.

Thirdly, we analysis the LOF technique, in this experiment, we computed the local outliers for a database of soccer-player information from the "Football 1. Bundesliga"(the German national soccer league) for the season 1998/99. The database consists of 375 players, containing the name, the number of games played, the number of goals scored and the position of the player (goalie, defense, center, offense). From these we derived the average number of goals scored per game, and performed outlier detection on the three-dimensional subspace of number of games, average number of goals per game and position (coded as an integer). In general, this dataset can be partitioned into four clusters corresponding to the positions of the players. We computed the *LOF* values in the *MinPts*

range of 30 to 50. Below we discuss all the local outliers with *LOF* > 1.5, and explain why they are exceptional.The strongest outlier is Michael Preetz, who played the maximum number of games and also scored the maximum number of goals, which made him the top scorer in the league. He was an outlier relative to the cluster of offensive players. The second strongest outlier is Michael Schjönberg. He played an average number of games, but he was an outlier because most other defense players had a much lower average number of goals scored per game. The reason for this is that he kicked the penal shots for his team. The player that was ranked third is Hans-Jörg But, a goalie who played the maximum number of games possible and scored 7 goals. He was the only goalie to score *any* goal; he too kicked the penalty shots for his team. On rank positions four and five, we found Ulf Kirsten and Giovane Elber, two offensive players with very high scoring averages.

Rank	Outlier Factor	Player Name	Games Played	Goals Scored	Position
1	1.87	Michael Preetz	34	23	Offense
2	1.70	Michael Schjönberg	15	6	Defense
3	1.67	Hans-Jörg Butt	34	7	Goalie
4	1.63	Ulf Kirsten	31	19	Offense
5	1.55	Giovane Elber	21	13	Offense
minimum			0	0	
median			21	1	
maximum			34	23	
mean			18.0	1.9	
standard deviation			11.0	3.0	

Table. 4: Results of the soccer player dataset [29]

Finally ,we analysis information theory based algorithms ,as we know this techniques only works upon unsupervised categorical datasets .In our experiment we conducted to see whether the solutions obtained by ITB-SS and ITB-SP are close to the optimal solutions obtained by optimizing the object function $j_x(y,o)$. The data set used is the public categorical "soybean data" [28], with 47 objects and 35 attributes. This data contains a very small class of 10 objects (numbers 11 to 20 in the original data set). Since the data does not have explicitly identified outliers, it is natural to treat the objects of the smallest class as "outliers". Therefore, we should check whether objects from this class will be detected for o = 1,....,10. Table 5 shows different sets of "outliers" obtained by ITB-SP, ITB-SS, and the optima for different values of o. The function $j_x(y,o)$ values in bold-faced letters indicate the cases where non-optimal sets were detected by either ITB-SP or ITB-SS, while the subsets of objects 11 to 20, which originally belong to the smallest class, found by strictly optimizing the function $j_x(y,o)$. are taken as reference sets of optimality. When the proces detect more outliers, the ITB is suits the optimization very well.

	ITB-SP	$J_x(y,o)$	ITB-SS	$J_x(y,o)$	Optimal	$J_x(y,o)$
1:	11	9.686	11	9.686	11	9.686
2:	11,18	9.687	11,18	9.687	11,18	9.687
3:	11,**15**,18	**9.687**	11,**15**,18	**9.687**	11,16,18	9.676
4:	11,15,16,18	9.671	11,15,16,18	9.671	11,15,16,18	9.671
5:	11,15,16,18,20	9.659	11,15,16,18,20	9.659	11,15,16,18,20	9.659
6:	11,15,**16**,18,19,20	**9.646**	11,13,15,18,19,20	9.642	11,13,15,18,19,20	9.642
7:	11,13,15,16,18,19,20	9.585	11,13,15,16,18,19,20	9.585	11,13,15,16,18,19,20	9.585
8:	11,13,**14**,15,16,18,19,20	**9.541**	11,13,15,16,17,18,19,20	9.537	11,13,15,16,17,18,19,20	9.537
9:	11,13,14,15,16,18,19,20,29	**9.493**	11,13,14,15,16,17,18,19,20	9.468	11,13,14,15,16,17,18,19,20	9.468
10:	11,12,13,14,15,16,18,19,20,29	9.419	11,12,13,14,15,16,17,18,19,20	9.334	11,12,13,14,15,16,17,18,19,20	9.334

Table. 5: Comparison among ITB-SP, ITB-SS, and Optimal Solutions on Soybean Data

VI. RELATED WORK

In the paper [3] the classic definition of an outlier is due to Hawkins who defines "an outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism".

In the paper[19] , most approaches on outlier mining in the early work are based on statistics which use a standard distribution to fit the dataset. Outliers are defined based on the probability distribution. For example, Yamanishi et a1. used a Gaussian mixture model to describe the normal behaviors and each object is given a score on the basis of changes in the model .

In this paper [20] Knorr et al. proposed a new definition based on the concept of distance, which regard a point p in data set as an outlier with respect to the parameters K and λ, if no more than k points in the data set are at a distance λ or less than p .

In this paper [21] Breunig et al. introduced the concept of local outlier, a kind of density-based outlier, which assigns each data a local outlier factor LOF of being an outlier depending on their neighborhood . The outlier factors can be computed very efficiently only if some multi-dimensional index structures such as R-tree and X-tree are employed. A top-n based local outlier mining algorithm which uses distance bound micro-cluster to estimate the density was presented in [22].

In the paper [18] the author has investigated outlier detection for categorical data sets. The problem is especially challenging because of the difficulty of defining a meaningful similarity measure for categorical data. The formal definition of outliers and an optimization model of outlier detection via a new concept of holoentropy that takes both entropy and total correlation into consideration. Based on this model two practical 1-parameter outlier detection methods named ITB-SS and ITB-SP which require no user-defined parameters for deciding whether an object is an outlier. The users need only provide the number of outliers they want to detect. The experimental results show that ITB-SS and ITB-SP are more effective and efficient than main stream methods and can be used to deal with both large and high-dimensional data sets where existing algorithms fail.

In the paper [23] they have introduced a distributed method for detecting distance-based outliers in very large data sets. Approach is based on the concept of outlier detection which is a small subset of the data set that can be also employed for predicting novel outliers. The used method exploits parallel computation in order to obtain vast time savings. Beyond preserving the correctness of the result the proposed schema exhibits excellent performances. Theoretically the cost of their algorithm is expected to be at least three orders of magnitude faster than the classical nested-loop outliers detector. Their experimental results show that the algorithm is efficient and that its running time scales well for an increasing number of nodes. The variant of the basic strategy which reduces the amount of data to be transferred in order to improve both the communication cost and the overall runtime.

In the paper [24] they have taken into consideration the class imbalance problem and offers new insights on similarity and redundancy of existing outlier detection

methods. As a result the design of effective ensemble methods for outlier detection is considerably enhanced.

VII. CONCLUSION

This paper mainly discusses about outlier detection approaches from data mining perspective. Firstly, we take overview of outlier ,types of outliers and outlier detection. Next, we reviews related work in outlier detection. Next, we discuss and compare algorithms of outlier detection which can be grouped into statistical-based approach, distance-based approach, density-based approach,Information theoretic-based approach. We discuss advantages and limitations of each algorithms. Finally, in implementation section , our experiments on different datasets show promising results, accurately finding outliers.

REFERENCES

[1] Yu, D., Sheikholeslami, G. and Zang, "A find out: finding outliers in very large datasets". In Knowledge and Information Systems, 2002, pp.387-412.

[2] Breunig, M.M., Kriegel, H.P., and Ng, R.T., "LOF: Identifying density-based local outliers." *ACM* Conference Proceedings, 2000, pp. 93-104.

[3] D. M. Hawkins, "Identification of Outliers". *Chapman and Hall*, London, 1980.

[4] Aggarwal, C. C., Yu, S. P., "An effective and efficient algorithm for high-dimensional outlier detection, The VLDB Journal, 2005, vol. 14, pp. 211-221.

[5] http://en.wikipedia.org/wiki/Outlier.

[6] Jiawai Han,Micheline Kamber,Jian Pei "Data Mining-concepts and techniques",morgan kaufmaan publishers,third edition.

[7] Hodge, V.J. and Austin, J. (2004) A survey of outlier detection methodologies. Artificial Intelligence Review, 22 (2). pp. 85-126.

[8] Barnett, V. & Lewis, T. (1994).,"Outliers in Statistical Data", 3rd edn. John Wiley & Sons.

[9] Rousseeuw, P. & Leroy, A. (1996).,"Robust Regression and Outlier Detection", 3rd edn. John Wiley & Sons.

[10] M. O. Mansur , Mohd. Noor Md. Sap,"Outlier Detection Technique in Data Mining: A Research Perspective", Proceedings of the Postgraduate Annual Research Seminar 2005,23-31.

[11] S. Ramaswamy,R. Rastogi,K.Shim"Efficient algorithms for mining outliers from large data sets". Proceedings of the International Conference on Management of Data, Dallas, Texas.2000.

[12] Z.Chen,a.Fu,J.Tang, "On Complementarity of Cluster and Outlier Detection Schemes".Springer Verlag,LNCS 2737, pp.234-243,2003 .

[13] Papadimitriou, S., Kitawaga, H., Gibbons, P., Faloutsos,C., "LOCI: Fast outlier detection using the local correlation integral", Proc. of the Int'l Conf. on Data Engineering,2003.

[14] Prasanta Gogoi, D K Bhattacharyya, B Borah and Jugal K Kalita, "A Survey of Outlier Detection Methods in Network Anomaly Identification", The Computer Journal, Received 27 September 2010; revised 9 February 2011.

[15] W. Lee and D. Xiang, "Information-Theoretic Measures for Anomaly Detection," Proc. IEEE Symp. Security and Privacy, 2001.

[16] Z. He, X. Xu, and S. Deng, "An Optimization Model for Outlier Detection in Categorical Data,"Proc. Int'l Conf. Advances in Intelligent Computing (ICIC '05), 2005.

[17] K. Das, J. Schneider, and D.B. Neill, "Anomaly Pattern Detection in Categorical Data Sets," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '08), 2008.

[18] Shu Wu, Shengrui Wang, "Information-Theoretic Outlier Detection for Large-Scale Categorical Data" IEEE Transactions On Knowledge and Data Engineering, VOL. 25, NO. 3, march 2013, pp589-602.

[19] Yamanishi. K, Takeuchi. J ,and Williams. G On-line, "unsupervised outlier detection using finite mixtures with discounting learning algorithms". In Proceedings of the Sixth ACM SIGKDDOO, Boston, MA, USA, pp.320-324.

[20] Knorr, E.M., Ng, R.T., "Finding Intentional Knowledge of Distance-Based Outliers", Proceedings of the 25th International Conference on Very Large Data Bases,Edinburgh, Scotland, pp.211-222, September 1999.

[21] Breunig, M.M., Kriegel, H.P., and Ng, R.T., "LOF:Identifying density-based local outliers." ACM Conference Proceedings, 2000, pp. 93-104.

[22] Jin, W., Tung, A.K.H., Han, J.W. "Mining Top-n Local Outliers in Large Databases". In: KDD (2001).

[23] 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, VOL. 25, NO. 7, JULY 2013.

[24] Erich Schubert Remigius Wojdanowski Arthur Zimek Hans-Peter Kriegel, "On Evaluation of Outlier Rankings and Outlier Scores" 12th SIAM International Conference on Data Mining (SDM), Anaheim, CA, 2012.

[25] SkyMark: Control Chart, at http://www.skymark.com/resources/tools/control_charts.asp (accessed: 13 December 2005)

[26] Wikipedia: Linear Regression, at http://en.wikipedia.org/wiki/Linear_regression (accessed: 13 December 2005)

[27] G.Williams, R. Baxter, H. He, S. Hawkins, L. Gu, "A comparative study of RNN for outlier detection in data mining". Proceedings of the 2nd IEEE International Conference on Data Mining (ICDM02) Maebashi City, Japan, 2002, pp. 709-712.

[28] UCI Machine Learning Repository, http://www.ics.uci.edu/ mlearn/MLRepository.html, 2011.

[29] Markus, Hans, Raymond,jorg, "LOF: Identifying density-based local outliers", Proc.ACM SIGMOD 2000 Int. Conf . On management of Data, Dalles,TX,2000.

[30] Zuriana, Rosmayat, Akbar , Mustafa ,"A Comparative Study for Outlier Detection Techniques in Data Mining", 1-4244-0023-6/06/$20.00 ©2006 IEEE .

