A Five Step Procedure for Outlier Analysis in Data Mining

Article · April 2012

CITATIONS
20

READS
9,106

European Journal of Scientific Research
ISSN 1450-216X Vol.75 No.3 (2012), pp. 327-339

http://www.europeanjournalofscientificresearch.com

A Five Step Procedure for Outlier Analysis in Data Mining

V. Ilango
Department of Computer Applications
New Horizon College of Engineering, Bangalore-560103, India
E-mail: banalysist@yahoo.com
Tel: +91-080-6629777

R. Subramanian
Department of Computer Science, Pondicherry University, Pondicherry, India
E-mail: rsmanian.csc@pondiuni.edu.in

V. Vasudevan
Department of Information Technology, Kalasalingam University, Srivilliputtur, India
E-mail: v.vasudevan@klu.ac.in

Abstract

Nowadays, outlier detection is primarily studied as an independent knowledge discovery process merely because outliers might be indicators of interesting events that have never been known before. Despite the advances seen, many issues of outlier detection are left open or not yet completely resolved. Outlier detection is an important data mining task. It deserves more attention from data mining community. There are "good" outliers that provide useful information that can lead to the discovery of new knowledge and "bad" outliers that include noisy data points. Distinguishing between different types of outliers is an important issue in many applications. It requires not only an understanding of the mathematical properties of data but also relevant knowledge in the domain context in which the outliers occur. We propose a novel five step procedure for outlier analysis along with a comprehensive review of existing outlier detection techniques. The paper ends by addressing some important issues and open questions that can be subject of future research. This paper would be helpful in devising the choice of outlier analysis techniques for unsupervised machine learning research.

Keywords: Univariate, Multivariate, Parametric, Nonparametric, detection rate, false alarm and ROC curve

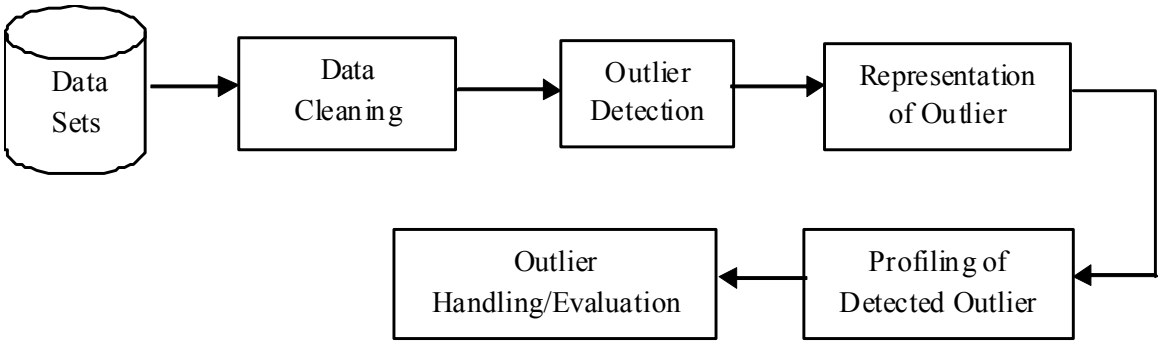
1. Introduction

Outliers are present in virtually every data set in any application domain, and the identification of outliers has a hundred years long history. Number definition are compiled and expressed in [104]. The important definitions are quoted here. "An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data" [7]. An outlying observation, or 'outlier', is one that appears to deviate markedly from other members of the sample in which it occurs [32]. An outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism [38]. Outlier detection methods have been suggested for numerous

applications, such as credit card fraud detection, clinical trials, voting irregularity analysis, data cleansing, network intrusion, severe weather prediction, geographic information systems, athlete performance analysis, and other data-mining tasks, fraud detection, medicine, public health, sports statistics, detecting measurement errors, loan application processing ,intrusion detection, activity monitoring , network performance ,fault diagnosis , structural defect detection , satellite image analysis , time-series data analysis , medical condition monitoring and pharmaceutical research [43]. Outliers may lead to the discovery of unexpected knowledge. In the 1880s when the English physicist Rayleigh measured nitrogen from different sources, he found that there were small discrepancies among the density measurements. After closer examination, he discovered that the density of nitrogen obtained from the atmosphere was always greater than the nitrogen derived from its chemical compounds by a small but definite margin. He reasoned from this anomaly that the aerial nitrogen must contain a small amount of a denser gas. This discovery eventually led to the successful isolation of the gas argon, for which he was awarded the Nobel Prize for Physics in 1904.which he considered as outlier that produced good outcome [30]. [9] [17] [23] [92] [105] [34] [60] provide an extensive survey of outlier detection techniques developed in machine learning and statistical domains. Our survey tries to provide a structured and comprehensive overview of the outlier analysis techniques. Outlier will arise due to natural variability of data set, measurement error as well as any recording error done by the users and execution error [45]. Robust estimation to find the presence of outliers in the given sample is a critical problem [100].

This study covers the answer for the following questions. a) How to define abnormality detection, b) How to minimize computational cost (processing time, storage and I/O) c) How to eliminate or minimize the impact of outlier in performance of information system, and to discover new knowledge from hidden data. The important issues associated with outliers are detecting the outlier and deciding what to do once it has been detected. Outlier detection involves identifying the time of occurrence, which may not be known, as well as recognizing the type of outlier [57]. A key challenge in outlier detection is that it involves exploring the unseen space. It is hard to enumerate all possible normal behaviors in an application. Handling noise in outlier detection is a challenge. Noise may distort the normal objects and blur the distinction between normal objects and outliers. It may help to hide outliers and reduce the effectiveness of outlier detection [87**].** The scope of this paper is modest to provide bird's eye view of outlier analysis techniques that focuses on unsupervised learning methods. The contributions of this paper are listed below: Section two describes the five step procedures for outlier analysis. Part (a) we have briefly explained different data sets and data cleaning measures. Part (b) broadly describes various techniques for outlier detection based on unsupervised approach. Part (c) discusses the methods of outlier representation. Part (d) narrates the techniques of how to profile and describe the detected outlier. Part (e) explores the interesting measures for evaluation of outlier. Finally paper concludes with suggestion for future research. To the best of our knowledge, this survey is an attempt to provide a structured and a comprehensive-overview of outlier analysis techniques

Figure 1: Five Step Procedure of Outlier Analysis



2. Outlier Analysis Procedure

We have proposed in figure.1 five step outlier analysis procedures starting from data sets, data cleaning, outlier detection, representation, profiling, handling and evaluation. Each step is explained in detail as follows. a). Data sets are important for outlier analysis. There are different types of data set such as: Nominal, ordinal, interval, ratio, binary, continuous, discrete, Transaction Data, Spatial Data, Spatio-Temporal Data, and Sequence Data and Time Series data [70]. *Data Cleaning:* Indentifying missing values is one of the data cleaning process. Missing values create difficulties for data analysis. The following measures can be used to process the missing values such as: Ignoring the record, can fill missing values manually, use global constant to fill in the missing values, use the attribute mean to fill in the missing values, use the attribute mean for all samples belonging to the same class as the given tuple [69] [76]. b*). Outlier Detection Techniques:* In the last decade numerous outlier detection methods have been proposed. The layout of outlier mining techniques has been explained in figure.2. The main focus is given on unsupervised outlier detection methods. Some of the outlier detection techniques can be used for generic purpose and some of them can be used for specific purpose [11] [41]. Outlier detection approaches can be classified into these three categories: *supervised, semi-supervised and unsupervised*. Techniques trained in supervised mode assume the availability of a training data set which has labeled instances for normal as well as anomaly class. Typical approach in such cases is to build a predictive model for normal vs. anomaly classes [66]. The unsupervised approach [86] of outlier detection does not require training data. This approach takes as input a set of unlabelled data and attempts to find outlier within the data. Many semi-supervised [10][85] techniques, assume that the training data has labeled instances for only the normal class, can be adapted to operate in an unsupervised mode by using a sample of the unlabeled data set as training data. Table 1 explains the advantages, drawback, techniques and tools for the above mentioned outlier detection methods. Unsupervised learning approaches can be further grouped into *ii).parametric and non-parametric methods* [28]. *Parametric method:* These methods assume that the whole data can be modeled to one standard statistical (normal) distribution. A point that deviates significantly from the data model is declared as an outlier. *Non-parametric method:* These methods make no assumption on the statistic properties of data and instead identify outliers based on the full dimensional distance measure between points.

Figure 2: Hierarchical Structure of Outlier Detection Methods

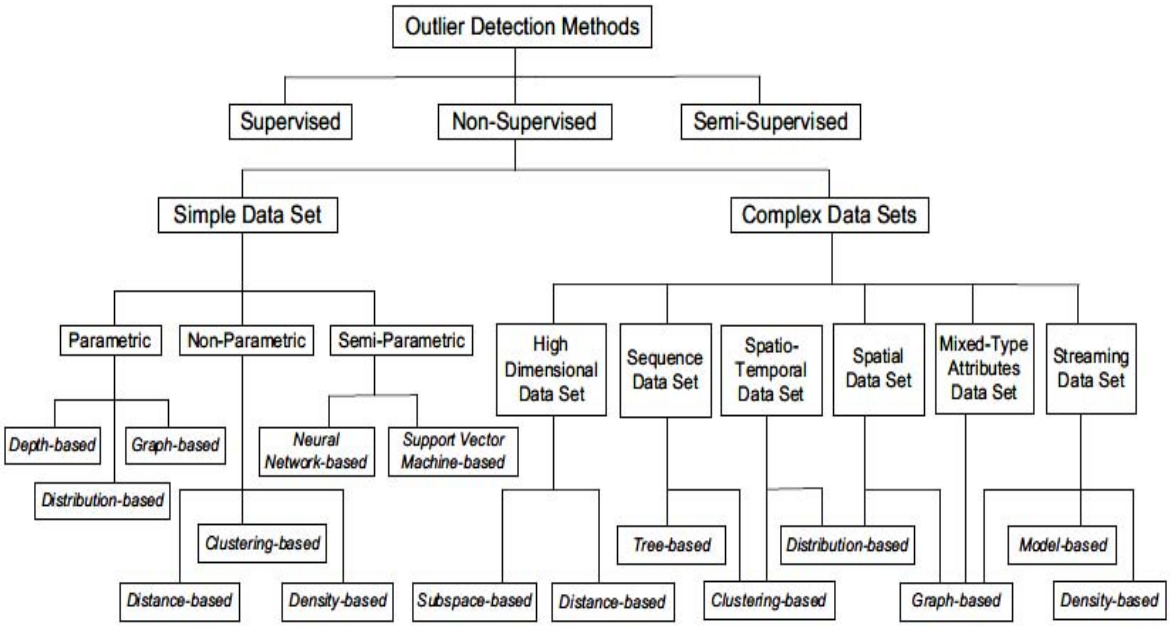


Table 1: Machine Learning Outlier Techniques

Supervised	**Semi-supervised**	**Unsupervised**
Require knowledge of both normal and anomaly class.Build classifier to distinguish between normal and known anomalies	Require knowledge of normal class only. Use modified classification model to learn the normal behavior and then detect any deviations from normal behavior as anomalous	Assume the normal objects are somewhat ``clustered'' into multiple groups, each having some distinct features .An outlier is expected to be far away from any groups of normal objects
Advantages: Models that can be easily understood.High accuracy in detecting many kinds of known anomalies Drawbacks:	Models that can be easily understood. Normal behavior can be accurately learned	The unsupervised techniques typically suffer from higher false alarm rate, because often times the underlying Assumptions do not hold true.
Require both labels from both normal and anomaly class. Cannot detect unknown and emerging anomalies	Require labels from normal class Possible high false alarm rate - previously unseen (yet legitimate) data records may be recognized as anomalies	Cannot detect collective outlier effectively. Normal objects may not share any strong patterns, but the collective outliers may share high similarity in a small area
Techniques: Artificial neural network, Bayesian statistics ,Rule based model, RBF,Ripper,SOM Decision tree learning	One class SVM,Hidden Markove Model	K-mean,EM, Wavecluster,CLAD,CBLOF,LOF,COF,DBSCAN,SNN
Software Tools :Weka ,SPSS,SAS,Tanagra, SYSTAT, MATLAB, Minitab, Liseral, MEDCALC,R-package		

Table 2: Comparison Of Outlier Detection Techniques For Simple Data Sets

Outlier Property				**Outlier Detection Technique Property**			**Data Sets Property**				**Reference**
Outlier Type		**Outlier Degree**		**Technique Based on**	**No of Outlier Detected at Once**		**Data Dimension**			**Data Type**	
Global	**Local**	**Scalar**	**Outlierness**		**One**	**Multiple**	**Univariate**	**Multivariate**			
								Moderate	**High**		
√		√		Distribution	√		√			Numeric	32
√		√		Distribution		√	√			Numeric	7
√		√		Distribution	√		√			Numeric	27
√			√	Distribution		√		√		Mixed-type	51
√		√		Depth		√		√		Numeric	80
√		√		Graph		√	√			Mixed-type	59
√		√		Clustering		√		√		Numeric	22
√			√	Clustering	√			√		Numeric	61
√	√		√	Clustering		√		√		Numeric	102
√			√	Clustering		√		√		Numeric	21
√		√		Distance		√		√		Numeric	52
√			√	Distance		√		√		Numeric	77
√			√	Distance		√		√		Numeric	8
	√		√	Density		√		√		Numeric	14
	√		√	Density		√		√		Numeric	20
	√		√	Density		√		√		Numeric	50
	√		√	Density		√		√		Numeric	96
	√		√	Density		√		√		Numeric	95
	√		√	Density		√		√		Numeric	83
	√		√	Density		√		√		Numeric	81
	√	√		Density		√		√		Numeric	78
	√		√	Density		√		√		Numeric	28

Table 2: Comparison Of Outlier Detection Techniques For Simple Data Sets - continued

	√		√	Density		√		√		Numeric	53
√			√	NN		√		√		Numeric	37
√		√		NN		√		√		Numeric	29
√		√		SVM		√		√		Numeric	63
√			√	SVM		√		√		Numeric	62

Table 5: Comparison of univariate, bivariate and multivariate data

	Univariate	**Bivariate**	**Multivariate**
Definition	Describes a case in terms of a single variable - the distribution of attributes that comprise it. Does not deal with causes or relationships	Analysis of two variables simultaneously. Focus is on the variables and the empirical relationships. Deals with causes or relationships.	Analysis of two or more variables simultaneously.
Graphical Representation	Bar graph, histogram, pie chart, line graph, box-and-whisker plot, Stem-and-leaf Plot, Q-Q Plots, violin plot	Scatter plot	Multivariate Profile, Andrew's Fourier Transformation, Chernoff's faces, Scatter plot, Contour Plots
Methods	Single and Sequential Procedure Inward and Outward Procedure Univariate Robust measure	-	Masking effect Swamping effect Multivariate Robust Measures
Computational Measures	Measure of Central Tendency Measures of dispersion Measures of Skewness One-way ANOVA Index Numbers Simple correlation & Regression	Simple Regression Simple Correlation Two-way ANOVA Association of attributes	Multiple regression & correlation Multiple discriminant analysis Multi-ANOVA Canonical analysis Factor analysis Cluster analysis & PCA

Outliers are considered as those points that are distant from their own neighbors in the data set. Compared to parametric methods, these non-parametric methods are more flexible and autonomous due to the fact that they require no data distribution knowledge [5][107]. Some of the computational measures for parametric and nonparametric test are discussed in table 4.

Table 4: Parametric and nonparametric test for outlier

Feature	**Parametric**	**Nonparametric**
Two samples – compare mean value for some variable of interest	t-test for independent samples	Wald-Wolfowitz runs test, Mann-Whitney U test, Kolmogorov-Smirnov two sample test
Multiple groups	Analysis of variance (ANOVA/ MANOVA)	Kruskal-Wallis analysis of ranks, Median test
Compare two variables measured in the same sample	t-test for dependent samples	Sign test, Wilcoxon's matched pairs test
If more than two variables are measured in same sample	Repeated measures ANOVA	Friedman's two way analysis of variance, Cochran Q
Two variables of interest are categorical	Correlation coefficient	Spearman R, Kendall Tau, Coefficient Gamma, Chi square, Phi coefficient, Fisher exact test, Kendall coefficient of concordance

Table 3: Comparison of outlier detection techniques for complex data sets

Outlier Property				Outlier Detection Technique Property			Data Sets Property				Reference
Outlier Type		Outlier Degree		Technique Based on	No of Outlier Detected at Once		Data Dimension			Data Type	
Global	Local	Scalar	Outlierness		One	Multiple	Univariate	Multivariate			
								Moderate	High		
√			√	Subspace		√			√	Numeric	3
√			√	Subspace		√			√	Numeric	106
√			√	Subspace		√			√	Numeric	65
√			√	Distance		√			√	Numeric	4
√			√	Distance		√			√	Numeric	31
√			√	Distance		√		√		Numeric	18
√			√	Graph		√		√		Mixed-type	25
	√		√	Graph		√		√		Mixed-type	58
	√		√	Graph	√			√		Mixed-type	39
√			√	Clustering		√		√		Sequence	15
√			√	Tree		√		√		Sequence	73
√		√		Distribution		√	√			Spatial	84
√		√		Distribution	√		√			Spatial	16
√			√	Distribution		√		√		Spatial	54
	√		√	Distribution		√		√		Spatial	94
	√	√		Distribution		√		√		Spatial	47
√			√	Model		√		√		Streams	102
√		√		Model		√		√		Streams	101
√		√		Graph		√		√		Streams	82
	√		√	Density		√		√		Streams	75
	√		√	Density		√		√		Streams	93
√		√		Clustering& Distribution		√		√		Spatial-temporal	19
√		√		Clustering& Distribution		√		√		Spatial-temporal	12

iii). Univariate, Bivariate and Multivariate Data sets: The technical description of different data set properties are described in Table 2, Table 3 and Table 5. Univariate analysis is the simplest form of quantitative (statistical) analysis. A basic way of presenting univariate data is to create a frequency distribution of the individual cases, which involves presenting the number of attributes of the variable studied for each case observed in the sample. This can be done in a table format, with a bar chart or a similar form of graphical representation.[44]. Bivariate data involves the analysis of two variables for determining the empirical relationship between them. Common forms of bivariate analysis involve creating a percentage table, a scatter plot graph, or the computation of a simple correlation coefficient [42]. Multivariate analysis (MVA) is based on the Analysis of two or more variables simultaneously [1][35]. Table 6 narrates the classification of outlier detection methods and their strength and weakness along with related algorithm[2][6][33][40][46][48][49][55][56][64][67][68][71][72][74] [88][89][90][98][99]. c).*Outlier Representation Stage:* Once outlier is detected, it must be represented in understandable form. The representation can be in the visual form of graphical display [26]. The goal of visualization is the interpretation of the visualized information. [108] explain the following principle for effective graphical display: Apprehension, Clarity, Consistency, Efficiency, Necessity and Truthfulness. [109] has also explained the following principle for graphical excellence: Graphical excellence consists of complex ideas communicated with clarity, precision and efficiency. Graphical excellence requires telling the truth about the data. Identified outlier can be represented using the above principles. d).*Profiling and Outlier Description:* Examples of outliers abound in social as well as scientific contexts. Outliers could also be indications of interesting events that have never been known before and hence; detecting outliers may lead to the discovery of critical information contained in data. In

such cases, uncovering underlying cause(s) is necessary. For example, if one happens to find an UFO (Unidentified Flying Object), throwing it away is obviously not a good idea. Studying its structure to understand its flying mechanism is certainly much more interesting and beneficial [79]. Once the outliers have been identified, the analyst should generate profile on each outlier observation and carefully examine the data for the variables responsible for its being an outlier. In addition to this the analyst can perform statistical and mathematical methods to identify the difference between outliers and the other observations. *Retention and Deletion:* After the outliers have been identified, profiled and categorized the analyst must decide on the retention or deletion of each one. Either deletion or accommodation of outlier depends on application domain, types of data sets and researchers. Person should have strong domain knowledge to decide about deletion and accommodation of detected outliers. If we want to reduce the weight of the outlier, we can use the following options: firstly if we have only a few outliers, we may simply delete those values, so they become blank or missing values. Secondly if there are too many outliers in a variable, or if we do not need that variable, we can delete the variable. Thirdly we can transform the values or variables. After dealing with the outlier, we re-run the outlier analysis procedure to determine if the data are outlier free. Sometimes new outliers emerge because they were masked by the old outliers and the data is now different after removing the old outlier so existing extreme data points may now qualify as outliers [110]. If new outliers emerge, and we want to reduce the influence of the outliers, we choose one of the above mentioned options again. Then, re-run the outlier analysis to determine if any new outliers emerge or if the data are outlier free, and repeat again [91]. e). *Evaluation of the Outlier:* Evaluation of detected outlier is an important task in the data analysis. It has number of measures; some of them are discussed as follow. *Detection Rate, False Alarm Rate and ROC Curves.* Intuitively, detection rate gives information about the number of correctly identified outliers, while the false alarm rate represents the number of outliers misclassified as normal data records. The most widely used tool to assess detection techniques' accuracy is ROC (Receiver Operating Characteristic) curve. *Computational Complexity.* The efficiency of outlier detection techniques can be evaluated by the computational cost, which is known as time & space complexity. In addition, the amount of memory occupation required to execute outlier detection techniques can be viewed as an important performance evaluation metrics. [24][36][60].

Table 6: Outlier Detection Approach

Approach	Definition	Strength	Weakness	Application	Methods/ Algorithm
Statistical Tests	Statistical methods assume that the normal data follow some statistical model. The data not following the model are outliers.	Utilize existing statistical modeling techniques to model various type of distributions	Most tests are for single attribute.With high dimensions, difficult to estimate distributions. Parametric assumptions often do not hold for real data sets	Fraud detection, Intrusion detection, Medical and health	parametric vs. non-parametric
Depth-based Approaches	Search for outliers at the border of the data space but independent of statistical distributions. Organize data objects in convex hull layers Outliers are objects on outer layers	Depth-based approaches avoid the problem of fitting to a data distribution. No assumption of probability distribution. No distance function required.	They are inefficient for the large data set with high dimensionality, where the convex hull will be harder to discern and is computationally more expensive.	Environmental monitoring, Localization and tracking, Logistics and transportation	ISODEPTH,FDC, Minimum Volume Ellipsoid (MVE) and Convex peeling.

Table 6: Outlier Detection Approach - continued

Distance-based Approaches	The concept of distance-based outlier relies on the notion of the neighborhood of a point, typically, the k-nearest neighbors.	This method avoids the excessive computation that can be associated with fitting the observed distribution into some standard distribution and in selecting discordancy tests.	Distance based method suffer from detecting local Outliers in a data set with diverse densities.	Intrusion detection, Environmental monitoring, Medical and Health care data	Index-based, Nested-loop based, Grid-based
Density-based Approaches	The density-based approach estimates the density distribution of the data and identifies outliers as those lying in low-density regions	Density-based techniques have the advantage that they can detect outliers that would be missed by techniques with a single, global criterion	Parameter selection for upper bound and lower bound is difficult.	Intrusion detection, Environmental monitoring, Medical and Health care, localization and tracking	Local outlier factor, k-distance, k-distance neighborhood, reach ability distance
Cluster Based Approaches	Cluster based approach finds groups of strongly related objects. An object is an outlier if it does not belong to any cluster, there is a large distance between the object and its closest cluster , or it belongs to a small or sparse cluster	Detect outliers without requiring any labeled data Work for many types of data. Clusters can be regarded as summaries of the data. Once the cluster are obtained, need only compare any object against the clusters to determine whether it is an outlier (fast)	Effectiveness depends highly on the clustering method used—they may not be optimized for outlier detection. High computational cost:	Fraud detection, Intrusion detection, Medical and health, Environmental monitoring, localization and tracking	BIRCH, CLARANS, DBSCAN, GDBSCAN, OPTICS and PROCLUS,CBLOF

3. Conclusions

In this paper we have proposed five step procedure of outlier analysis and tried to provide a broad view of latest techniques associated with each steps but obviously, we are unable to describe all approaches in a single paper. The limitation of this study is focused mainly on outlier detection techniques for low-dimensional simple static data, followed by some of recent advancements in outlier detection for high-dimensional data. Based on our review, we observe that the notion of outlier is different for different application domains. Outlier detection is an extremely important problem. It involves exploring unseen spaces. Some outlier detection techniques are developed in a more generic fashion and can be ported to various application domains while others directly target a particular application domain. Outlier detection is an important research problem in data mining that aims to discover useful abnormal and irregular patterns hidden in large data sets. Most existing outlier detection methods only deal with static data with relatively low dimensionality. Recently, outlier detection for high-dimensional stream data, ensemble outlier detection, and subspace outlier mining, addressing the issues of concept drift, dimension reduction, and detection result visualization became a new emerging research problem. Minimum number of research has been done using categorical data. Robust methods to be discovered

to explore interesting patterns. This necessitates the development of relevant approaches to handle the issue. These are to point out that outlier detection is a very active field of data mining research and an extensive study will bring many benefits to various practical applications as mentioned above.

References

[1] A.C. Atkinson, et.al, "Exploring Multivariate Data with the Forward Search". Springer-Verlag – New York. (2004).

[2] Agarwal, D., "Detecting anomalies in cross-classified streams: a Bayesian approach," *Knowl. Inf. Syst.*, vol. 11, no. 1, pp. 29–44, 2006.

[3] Aggarwal and P. S. Yu. An e®ective and e±cient algorithm for high-dimensional outlier detection. *International Journal on Very Large Data Bases*, 14(2):211{221, 2005.

[4] Angiulli and C. Pizzuti. Outlier mining in large high-dimensional data sets. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):203{215, 2005.

[5] Angela Hebel "Parametric vs Nonparametric Statistics- when to use them and which is more powerful?" university of Maryland, April, 2002.

[6] Arning, R. Agrawal, and P. Raghavan, "A Linear Method for Deviation Detection in Large Databases," Proc.Int' Conf. Knowledge Discovery and Data Mining, 1996, pp. 164-169.

[7] Barnett V, Lewis T. *Outliers in Statistical Data*. New York, NY: John Wiley & Sons; 1994.

[8] Bay and M. Schwabacher (2003) Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In: Proceedings of ACM SIGKDD, pp 29-38.

[9] Ben-Gal I. Outlier detection. In: Maimon O, Rockach L, eds. Data Mining and Data Discovery Handbook: A Complete Guidance for Practitioners and Researchers. Springer, US; 2005, I:131–146.

[10] Bennett K, Demiriz A (1998) Semi-supervised support vector machines. Adv Neural Inf Process Syst 12:368–374.

[11] Bhuyan, M. H., Bhattacharyya, D. K., and Kalita,J. K. (2011) Rodd: An effective reference based outlier detection technique for large datasets. *LNCS-CCIS*, 133, Part I, 76–84.

[12] Birant, A. Kut (2006) Spatio-temporal outlier detection in large database", In: Proceedings of ITI.

[13] Borgwardt, K. M., Ong, C. S., Sch¨onauer, S., Vishwanathan, S.V. N., Smola, A. J., and Kriegel, H.-P., "Protein function prediction via graph kernels," *Bioinformatics*, vol. 21, no. 1, pp. 47–56, 2005.

[14] Breunig MM, Kriegel H-P, Ng RT, Sander J (2000) LOF: identifying density-based local outliers. In: Proceedings of ACM SIGMOD, pp 93-104.

[15] Budalakoti, S. Cruz, A. N. Srivastava, R. Akella, E. Turkov (2006) Anomaly detection in large sets of high-dimensional symbol sequences. NASA TM.

[16] C.T. Lu, D. Chen, and Y. Kou (2003) Detecting spatial outliers with multiple attributes.In: Proceedings of ICTAI, pp 122-128.

[17] Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey, *ACM Comput Surv* 2009, 41, Article 15:1–58.

[18] Chaudhary, A. S. Szalay, and A. W. Moore (2002) Very fast outlier detection in large multidimensional data sets. In: Proceedings of ACM SIGMOD Workshop on DMKD.

[19] Cheng and Z. Li (2006) A multiscale approach for spatio-temporal outlier detection. Transactions in GIS, vol. 10, no. 2, pp 253-263.

[20] Chiu, A. W. Fu (2003) Enhancements on local outlier detection. In: Proceedings of IDEAS.

[21] D. Ren, I. Rahal, W. Perrizo (2004) A vertical outlier detection algorithm with clusters as by-product. In: Proceedings of ICTAI.

[22] D. Yu, G. Sheikholeslami, and A. Zhang (2002) Findout: ¯nding outliers in very large datasets. In: Journal of Knowledge and Information Systems, vol. 4, no. 3, pp. 387-412.

[23] Dav d Tan ar, "Research and Trends in Data Mining Technologies and Applications", Monash Un vers ty, Australia, Idea Group Publishing,2007.
[24] E Achtert, H.-P. Kriegel, L. Reichert, E. Schubert,R. Wojdanowski, and A. Zimek. Visual evaluation of outlier detection models. In *Proc. DASFAA*, 2010.
[25] E. Otey, A. Ghoting, S. Parthasarathy (2006) Fast distributed outlier detection in mixed-attribute data sets. Data Mining and Knowledge Discovery, vol. 12, no. 2-3, pp 203-228.
[26] Eberle, W. and Holder, L., "Anomaly detection in data represented as graphs," *Intell. Data Anal.*, vol. 11, no. 6, pp. 663–689, 2007.
[27] Eskin (2000) Anomaly detection over noisy data using learned probability distributions.In: Proceedings of Machine Learning.
[28] Fan, H., Za¨ıane, O. R., Foss, A., and Wu, J., "A nonparametric outlier detection for effectively discovering top-n outliers from engineering data," in *PAKDD*, pp. 557–566, 2006.
[29] Fu. J, X. Yu (2006) Rotorcraft acoustic noise estimation and outlier detection. In: Proceedings of IJCNN, pp 4401-4405.
[30] Gongxian Cheng, *"Outlier Management In Intelligent Data Analysis"*, Thesis report, Birkbeck College, University of London, 2000, p.17.
[31] Ghoting, S. Parthasarathy, and M. Otey. Fast mining of distance-based outliers in high-dimensional datasets. *Data Mining and Knowledge Discovery*, 16(3):349–364, June 2008.
[32] Grubbs, Frank (1969) Procedures for detecting outlying observations in samples. Technometrics, vol. 11, no. 1, pp. 1-21.
[33] H. V. Nguyen and V. Gopalkrishnan. Feature extraction for outlier detection in high-dimensional spaces. In *Proceedings of the 4th International Workshop on Feature Selection in Data Mining*, pages 64-73, 2010.
[34] Hadi AS, Imon A, Werner M. Detection of outliers.*Wiley Interdiscip Rev Comput Stat* 2009, 1:57–70.
[35] Hair. et.al , " Multivariate data analysis", Pearson Education Pte Ltd, Fifth edition, ISBN:81-297-0021-2.
[36] Hans-Peter Kriegel Peer Kr¨oger Erich Schubert Arthur Zimek, "Interpreting and Unifying Outlier Scores", 11th SIAM International Conference on Data Mining (SDM), Mesa, AZ, 2011.
[37] Harkins, H. He, G. J. Willams, R. A. Baster (2002) Outlier detection using replicator neural networks. In: Proceedings of DaWaK, pp 170-180.
[38] Hawkins DM. Identification of outliers. New York, NY: Chapman and Hall; 1980.
[39] He, S. Deng, X. Xu (2005) An optimization model for outlier detection in categorical data. In: Proceedings of ICIC, pp 400-409.
[40] Hewahi, N. Saad, M.: *Class Outliers Mining: Distance Based-Approach*, International Journal of Intelligent Systems and Technologies, Vol. 2, No. 1, pp 55-68, 2007.
[41] High Wycombe, and Buckinghamshire, Missing Values, Outliers, Robust Statistics & Non-parametric Methods, Shaun Burke, RHM Technology Ltd, , UK.
[42] Ho, Robert, "Handbook of univariate and multivariate data analysis and interpretation with SPSS", ISBN 1-58488-602-1, 2006 by Taylor & Francis Group.
[43] Hodge VJ, Austin J. A survey of outlier detection methodologies. *Artif Intell Rev* 2004, 22:85–126.
[44] http://en.wikipedia.org/wiki/Univariate_analysis
[45] Huang, H.Y., Lin, J.X., Chen, C.C and Fan, M.H.: Review of Outlier Detection. In: Application Research of Computers, 8–13. (2006).
[46] J. X. Yu, W. Qian, H. Lu, A. Zhou (2006) Finding centric local outliers in categorical/numerical spaces. Knowledge Information System, vol. 9, no. 3, pp 309-338.
[47] J. Zhao, C.-T. Lu, and Y. Kou (2003) Detecting region outliers in meteorological data. In: proceedings of ACM GIS, pp 49-55.
[48] Janeja VP, Atluri V. Spatial outlier detection in heterogeneous neighborhoods. *Intell Data Anal* 2009, 13:85–

[49] Jiangab, F., Suia, Y., and Caoa, C. (2008) A rough set approach to outlier detection. *International Journal of General Systems*, 37, 519–536.
[50] Jin, A.K.H. Tung, and J. Han (2001) Mining top-n local outliers in large databases. In: Proceedings of ACM SIGKDD, pp. 293-298.
[51] K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne (2000) On-line unsupervised learning outlier detection using finite mixtures with discounting learning algorithms. In: Proceedings of KDD, pp 250-254.
[52] Knorr E, Ng R(1998) Algorithms for mining distance-based outliers in large data sets. In:Proceedings of VLDB, pp 392-403.
[53] Kollios, D. Gunopulos, N. Koudas, S. Berchtold (2003) Efficient biased sampling for approximate clustering and outlier detection in large data sets. Knowledge and Data Engineering, vol. 15, no. 5, pp 1170-1187.
[54] Kou, C. Lu, D. Chen (2006) Spatial weighted outlier detection. In: Proceeding of SDM.
[55] Koufakou, M. Georgiopoulos, and G. Anagnostopoulos. Detecting outliers in high-dimensional datasets with mixed attributes. In *International Conference on Data Mining (DMIN 2008)*, L.Vegas,NV, 14-17 2008.
[56] Kriegel, H.-P., hubert, M. S., and Zimek, A., "Angle-based outlier detection in high-dimensional data," in *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York, NY, USA), pp. 444–452, ACM, 2008.
[57] Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A., " Outlier detection techniques", In Proc of SIAM International Conference on Data Mining, 2010.
[58] L. Wei, W. Qian, A. Zhou, W. Jin, J. X. Yu (2003) HOT: hypergraph-based outlier test for categorical data. In: Proceedings of PAKDD, pp 399-410.
[59] Laurikkala, M. Juhola, E. Kentala (2000) Informal identification of outliers in medical data. In: Proceedings of IDAMAP.
[60] Lazarevic, A., Ozgur, A., Ertoz, L., Srivastava, J. and Kumar, V. (2003) 'A comparative study of anomaly detection schemes in network intrusion detection', SIAM Conference on Data Mining.
[61] M. F. Jiang, S. S. Tseng, C. M. Su (2001) Tw-phase clustering process for outliers detection. Pattern Recognition Letters, 22 (6-7): 691-700.
[62] M. I. Petroveskiy (2003) Outlier detection algorithms in data mining system. Programming and Computer Software, vol. 29, no. 4, pp 228-237.
[63] M. J. Tax and R. P. W. Duin (1999) Support vector domain description. Pattern Recognition Letters, vol. 20, pp 1191-1199.
[64] M. Knorr, R. T. Ng, and V. Tucakov. Distance-based outliers: Algorithms and appli-cations. *International Journal on Very Large Data Bases*, 8(3-4):237, 2000.
[65] M. Shyu, S. Chen, K. Sarinnapakorn, L. W. Chang (2003) A novel anomaly detection scheme based on principal component classi¯er. In: Proceedings of ICDM, pp172-179.
[66] M. V. Joshi, R. C. Agarwal, and V. Kumar, "Mining needle in a haystack: classifying rare classes via two-phase rule induction," in Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, 2001, pp. 293–298.
[67] M¨orchen, F., "Unsupervised pattern mining from symbolic temporal data," *SIGKDD Explor. Newsl.*, vol. 9, no. 1, pp. 41–55, 2007.
[68] McQuarrie A, Tsai CL. Outlier detections in AR models. *J Comput Graph Stat* 2003, 12:450–471.
[69] Micheline Kamber and Jiawei Han, *Data Mining: Concepts and Techniques*.Morgan Kaufmann Publishers, 2 ed., Mar. 2006.
[70] N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Addison-Wesley, 2005.
[71] Ng, B. (2006) Survey of anomaly detection methods. Technical Report UCRL-TR-225264. Lawrence Livermore National Laboratory, University of California, California USA.

[72] NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook
[73] P. Sun, S. Chawla, B. Arunasalam (2006) Mining for outliers in sequential databases. In: Proceedings of SIAM, pp 94-105.
[74] Papdimitriou, P., Dasdan, A., and Garcia-Molina, H., "Web Graph Similarity for Anomaly Detection," technical report, Stanford, 22 Jan. 2008.
[75] Pokrajac, A. Lazarevic, L. J. Latechi (2007) Incremental local outlier detection for data streams. In: Proceedings of CIDM.
[76] R.J.A. and Rubin D.B. "Statistical Analysis with missing data", Wiley-Interscience and related papers, (2002).
[77] Ramaswamy S, Rastogi R, Shim K (2000) efficient algorithms for mining outliers from large data sets. In: Proceedings of ACM SIGMOD, pp 427-438.
[78] Ren, B. Wang, W. Perrizo (2004) RDF: a density-based outlier detection method using vertical data representation. In: Proceedings of ICDM, pp 503-506.
[79] Nguyen Hoang Vu, "Outlier Detection Based on Neighborhood Proximity", Dissertation report, Nanyang Technological University, June, 2010, p.1.
[80] Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection, 3rd edn.* John Wiley and Sons, 1996.
[81] S. Kim, S. Cho (2006) Prototype based outlier detection. In: Proceedings of IJCNN, pp 820-826.
[82] S. Muthukrishnan, R. Shah, J. S. Vitter (2004) Mining deviants in time series data streams. In: Proceedings of SSDBM.
[83] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. (2003) LOCI: fast outlier detection using the local correlation integral. In: Proceedings of ICDE, pp 315-326.
[84] S. Shekhar, C.-T. Lu, and P. Zhang (2001) A uni¯ed approach to spatial outliers detection.GeoInformatica, 7(2): 139-166.
[85] S. Y. Jiang, X. Song, H. Wang, J. J. Han, and Q. H. Li. A clustering-based method for unsupervised intrusion detections. *Pattern Recognition Letters*, 27:802–810, 2006.
[86] S. Zanero and S. M. Savaresi, "Unsupervised learning techniques for an intrusion detection system," in Proceedings of the 2004 ACM symposium on Applied computing, 2004, pp. 412 – 419.
[87] Sabyasachi Basu · Martin Meckesheimer "Automatic outlier detection for time series: an application to sensor data-Knowl Inf Syst (2007) 11(2): 137–154, DOI 10.1007/s10115-006-0026-6.
[88] Salvador, S. and Chan, P., "Learning states and rules for detecting anomalies in time series," *Applied Intelligence*, vol. 23, no. 3, pp. 241–255, 2005.
[89] Serneels, S. and Verdonck, T., "Principal component regression for data containing outliers and missing elements," *Comput. Stat. Data Anal.*, vol. 53, no. 11, pp. 3855–3863, 2009.
[90] Shaft, U. and Ramakrishnan, R., "Theory of nearest neighbors indexability," *ACM Trans. Database Syst.*, vol. 31, no. 3, pp. 814–838, 2006.
[91] http://www.psychwiki.com/images/7/79/lab1datascreening.doc
[92] Steinwart, I., Hush, D., and Scovel, C., "A classification framework for anomaly detection," *J. Mach. Learn. Res.*, vol. 6, pp. 211–232, 2005.
[93] Subramaniam, S., Palpanas, T., Papadopoulos, D., Kalogeraki, V., and Gunopulos, D., "Online outlier detection in sensor data using nonparametric models," in *VLDB '06: Proceedings of the 32nd international conference on Very large data bases*, pp. 187–198, VLDB Endowment, 2006.
[94] Sun, S. Chawla (2004) On local spatial outliers. In: Proceedings of ICDM, pp 209-216.
[95] T. Hu, S. Y. Sung (2003) Detecting pattern-based outliers. Pattern Recognition Letters, 24 (16): 3059-3068.

[96] Tang, J., Chen, Z., Fu, A. W., and Cheung, D. W.(2006) Capabilities of outlier detection schemes in large datasets, framework and methodologies. *Knowledge and Information Systems*, 11, 45–84.

[97] Vinueza, A. and Grudic, G. 2004. Unsupervised outlier detection and semi-supervised learning. Tech. Rep. CU-CS-976-04, Univ. of Colorado at Boulder. May.

[98] Witten and Frank, " Data Mining-Practical Machine learning Tools and Techniques", Morgan Kaufmann Publishers, Second Edition, 2005, ISBN: 0-12-088407-0.

[99] Y. Tao, X. Xiao, and S. Zhou. Mining distance-based outliers from large databases in any metric space. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 394-403, 2006.

[100] Walter R. Mebane and Jasjeet S. Sekhon, "Robust Estimation and Outlier Detection for over dispersed Multinomial Models of Count Data", *American Journal of Political Science*, Vol. 48, No. 2, April 2004, Pp. 392–411.

[101] Yamanishi, J. Takeuchi (2006) A unifying framework for detecting outliers and change points from non-stationary time series data. Knowledge and Data Enginnering, vol. 18, no. 4, pp 482-492.

[102] Z. He, X. Xu, S. Deng (2003) Discovering cluster based local outliers. Pattern Recognition Letters, 24 (9-10): 1651-1660.

[103] Zhang, K., Shi, S., Gao, H. and Li, J. (2007) 'Unsupervised outlier detection in sensor networks using aggregation tree', Proceedings of ADMA.

[104] Zhang, Y., Meratnia, N. and Havinga, P. J. M. (2007) 'A taxonomy framework for unsupervised outlier detection techniques for multi-type data sets', Technical Report, University of Twente.

[105] Zhang, Y., Meratnia, N., and Havinga, P. (2010) Outlier detection techniques for wireless sensor networks:A survey. *IEEE Communications Survey & Tutorials*,12(2), 159 – 170.

[106] Zhu, H. Kitagawa, C. Faloutsos (2005) Example-based robust outlier detection in high dimensional datasets. In: Proceedings of ICDM, pp 829-832.

[107] http://www.biomedcentral.com/1471-2288/5/35

[108] http://www.datavis.ca/gallery/accent.php.

[109] http://idt.stanford.edu/idt1999/students/mzuno/portfolio/work/reports/tufte

[110] http://216.22.10.76/wiki/Dealing_with_Outliers