ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

Surfeit Entropy-Based Outlier Detection for High-Dimensional Categorical Data Set

Neha L. Bagal, Prof. Y.B.Gurav

Department of Computer Engineering, PVPIT, Bavdhan, Pune, India

ABSTRACT: Classification, clustering, frequent patterns and statistics are different methods used for extraction of required information by removing unwanted data. Outlier detection from unsupervised data set is more challenging task as there is no way to measure the distance between objects. In this work we have proposed novel framework for outlier detection based on information theoretic measure using surfeit entropy. In this project we used information theoretic measures such as entropy and dual correlation. Using this model we proposed outlier detection algorithm and discussed results of proposed system.

KEYWORDS: Outlier detection, surfeit entropy, attributes weighting, dual correlation.

I. INTRODUCTION

Many researchers have concentrated on outlier detection problem in different domains and application areas. The outlier detection method detects distinct and exceptional data in majority of given input data.

Outlier detection is an important research problem which is researched within various research areas and application domains. Outlier detection method detects distinct, exceptional and inconsistent objects with respect to the majority data in a given input data sets. Number of outlier detection techniques has beenspecifically developed for some application areas [1].

Outlier detection is used in scientific research work for analyzing data and knowledge discovery process in astronomy, chemistry, biology etc.Outlier arises due to faults in mechanical systems, behavioral changes, and fraudulentnature [1].

The outlier detection helps to identify the systems faults before they affect to the outcomes of system.The techniques or algorithms used in outlier detection methods are varied notably which are mainly dependent upon the characteristics of data sets to be worked with [2].

Wecan classify the existing methods for outlier detection according to the availability of labels in the training data set;Following are three broad categories [1]:

1. Supervised,
2. Semi-supervised, and
3. Unsupervised approaches.

In supervised and semi-supervised approaches training needs to given before use.While the unsupervised approach do not include the training before use. For a supervised approach a training set should be provided with labels for anomalies as well as labels of outliers, in semi-supervised approach the training set with normal object labels alone required. Unsupervised approach does not requireany object label information.

These three approaches have different pre requirements, limitations and uses different data sets with different amounts of label information.All of these are discussed[8] in detail below.

Supervised outlier detection approach uses labeledobjects belonging to the normal and outlier classes tolearn the classifier and assign appropriate labels totest objects [2].

DOI: 10.15680/ijircce.2015.0306165

Semi-supervised outlier detection approach firstly learns normal behavior from given training data set of normal objects and thencalculates the similarity of test objects [2].

Unsupervised outlier detection approach detects outliers in unlabeled data set [1]. Assumes that the most of the objects in data set are normal. This approach is applied to various kinds of outlier detection methods and data sets [2]. In this paper we used the unsupervised method. To use supervised and semi-supervised approach one must first label the training data sets [2].When we considerlarge data sets or high dimensional data then labeling will be tedious and time consuming task [1].

A. Objectives

1) Outliers are the patterns that do not confirm to expected normal behavior [1]. The patterns which does not confirms to the normal behavior of objects are detected as outliers. But there are many factors which makes this simple task very challenging.
2) Defining a boundary between normal and anomalous behavior is not a simple task, but few anomalous object may appear like normal which makes outlier detection problem more complex and difficult.
3) Unsupervised method are applicable only on numerical data sets, however they cannot be used to deal with categorical data [2].
4) Using formal definition of outlier our aim is to develop effective and efficient method that can be used to detect outliers in large scale categoricalunsupervised data sets.
5) We have combined entropy and dual correlation with attribute weighting resulting intoweighted surfeit entropy where entropy computes uncertainty and dual correlation measures mutual information or attribute relation [2].

II. RELATED WORK

Proximity based method is used to measure compactness of objects in terms of distance / density [1]. CNB [11] and ORCA [10] are different algorithms for outlier detection in categorical data. ORCA uses hamming distance and CNB usescommon neighbor set [10]. Bothmethods are not useful for high dimensional data due to theirdifficulty in choosing the distance or density as well as high time and space complexity.

Rule-based methods use the concept of frequent items from association-rule mining [12]. Rule based method assumes the frequent or infrequent items as a data set.Objects with few frequent items or many infrequent items are more likely to be considered as outlier objects than others.

Otey’s algorithm [13] and frequent pattern outlier factor [12] are twowell known ruled based techniques.FIB algorithm includes aninitial computation of the set of frequent patterns, using apredefined minimum support rate. Allsupport rates of associated frequent patterns are summedup for each object as the outlier factor of this object [12]. While Otey’s algorithm, begins with computation of infrequent items from data set.Outlier factor is calculated using the same. Objects with largest scores are treated as outliers.

Random walk, Hyper-graph theory[8] methods are implemented using several approaches. Random walk method outliers are the object who has the low probability to combine with neighbor [14].Inmethod [15] relationships are considered and mutualdependence based local outlier factor is proposed todetect outliers. In other methods clusterbased local outlier detection method, classificationbased method.

Most of the existing systems are depends on user defined parameters and very few methods are dealing with unsupervised categorical data [7]. Therefore there is a need of a method which will be able to deal large scale categorical data without requirement ofany user defined parameter [2]. Also there is requirement of method which will perform outlier detection using joint correlation between attributes.

DOI: 10.15680/ijircce.2015.0306165

III. PROBLEM FORMULATIONS

In this section we first look at how entropy and dual total correlation can be used to capture similarity between outlier candidates. We are proposing WeightedSurfeit Entropy and formulate the outlier detection problem as in [2].

A. Entropy: Entropy is measure of informationand uncertainty of a random variable.

Let X be the set of n objects {x1,x 2,x 3 , , ,.... n x} , each xi for 1≤I ≤ n being a vector of categorical attributes [y 1,y 2,y 3 , , ,....ym]T where m number of isattributes. Now based on chain rule of entropy [2], Entropy of y denoted as Hx(y) can be written as follows.

$$Hx(y) = Hx(y1, y2, \dots, ym)$$

$$= \sum_{i=1}^{m} Hx(yi|yi-1, \dots . y1)$$

$$= Hx(y1) + Hx(y2|y1) + \dots + Hx(ym|ym-1, \dots, y1) \qquad (4)$$

Where

$$Hx(ym|ym-1, \dots, y1) = -\sum_{ym,ym-1,\dots,y1} p(ym, ym-1, \dots y1) \log p(ym|ym-1, \dots y1). \ (1)$$

Entropy of dataset decreases significantly with removal good outlier candidates.

B. Total Correlation:

It is defined as summation of mutual information of multivariate discrete random vector y, [2]and it is denoted as Cx(y). Total correlation is based on Watanabe's proof. Totalcorrelation can be expressed as :

$$Cx(y) = \sum_{i=1}^{m} Hx(yi) - Hx(y) \quad (2)$$

C. Dual total correlation:

The dual total correlation [2] calculates the amount of entropy present in Y beyond the sum of the entropies for each variable conditioned upon all other variables. The dual total correlation is also called as the surfeit entropy and the binding information. In this paper dual total correlation as

Sx(Y) and expressed as

$$Sx(y) \equiv (\textstyle\sum_{Xi\in X} Hx\backslash xi(y)) - (n-1)Hx(y) (3)$$

Where n is number of attributes.

To weight the entropy of each attribute, we are usinga reverse function of the entropy, as follows:

$$Wx(yi) = 2\,(1 - \frac{1}{1+\exp\,(-Hx(yi))} \ (4)$$

The weighted Surfeit entropy is defined as follows:

Definition 1: The weighted surfeit entropy EWX (Y)is the sum of weighted entropy on each attribute ofthe random vector Y.

$$SWx(Y) = \sum_{i=1}^{m} Wx(yi)Hx(yi) \qquad (5)$$

Outliers are detected by minimizing the surfeitentropy through the removal of outlier candidates;proposed strategyHave weighting the entropy of each individualattribute in order to give more importance to thoseattributes with small entropy values.

D. Formal Definition of outlier detection:

We are using weighted surfeit entropy for outlier detection outliers .We consider that set ofoutlier candidates is the best if entropy of dataset significantly decreases with its removal from dataset.

Definition 2: X be a given dataset with n objects and asubset Out(o) is defined as the set of outliers if itminimizes the weighted surfeit entropy of dataset X with oobjects removed.

E. Differential Entropy:

Definition 3: Given an object xo of X, the differenceof weighted surfeit entropy ex(xo) between the dataset X and the data set X\{xo} is defined as thedifferential surfeit entropy of the object xo.

$$sx(xo) = Wx(y) - Wx\{xo\}(y) \qquad (6)$$

F. Outlier Factor:

Outlier factor is a measure of how likely xois an outlier. An object xowith a largeoutlier factor value is more likely to be an outlier thanan object with a small value. Outlier factor of anobject xois denoted as OF(xo) is defined as:

$$OF(xo) = \sum_{i=1}^{m} OF(xo, i) \qquad (7)$$

$$OF(xo) = \sum_{i=1}^{m} \begin{cases} 0 & , \text{if } n(xo, i) = 1; \\ SWx(yi).\, \delta[n(xo, i)] & , \text{else}. \end{cases}$$

IV. PROPOSED APPROACH

Our proposed approach is based on weighted entropy anddifferential entropy which can be calculated usingequation (8) and (9). System will take data set file offormat .CSV and gives output file with outliers removed.

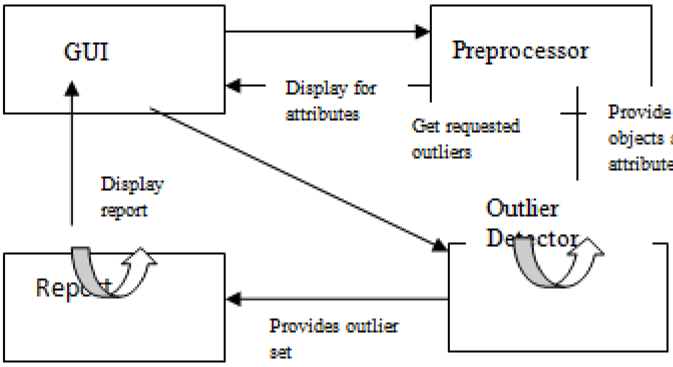


Fig. 1. System Architecture

A.System Architecture

To address the problem discussed above inneed of effective outlier detection in unsuperviseddata set. A proposed methodology with surfeitentropy and to deal with large scale categorical datais considered as shown in Fig. 1

B.Workflow of system

1. GUI takes input data file of .CSV format and gives file to preprocessor.
2. Preprocessor separates objects and attributes from data file and gives set of attributes to GUI to further display it to user.
3. Preprocessor gives objects and attributes to outlier detector component for further processing.
4. GUI sends request for detecting outliers to outlier detector component.
5. Outlier detector component detects outliers from given data set by using entropy, total correlation and surfeit entropy.
6. Outlier detector component removes outlier candidates and sends data set file to report generator.

DOI: 10.15680/ijircce.2015.0306165

7. Report generator component generates report and comparison model using graph. Report generator finally displays report via GUI.

C. Mathematical Model

Let Sbe the system representing outlier detection system, I be a set of inputs to the system and O represents the set of outputs generated by the system.

S= { I, O }

S= {Xi, Hx(Y),Wx(Y), SWx(Y), OF(xo)}

Where,

I = {Xi, Hx(Y), Wx(Y), SWx(Y)}

D. ModifiedAlgorithm for outlier detection

Here, we have derived Surfeit entropy basedsingle pass greedy algorithm for outlier detection. Outlier factors are computed only once,and the o objects with largest values are identified asoutliers [2]. This algorithm is parameter-less as we donot need to provide any user defined parameters.

```
Proposed Algorithm:
Input: Data Set X
Output: Outlier Set O
  1. Compute entropy of each object of random vector Y by using equation
     $Hx(Y) = Hx(y1) + Hx(y2|y1) + \dots + Hx(ym|ym-1, \dots, y1)$
for 1≤i≤m
  2. Compute total correlation Cx(Y) for random vector Y using equation $Cx(y) = \sum_{i=1}^{m} Hx(yi) - Hx(y)$
     For 1≤i≤m
  3. Compute weighted surfeit entropy for each attribute of random vector Y using equation
     $SWx(Y) = \sum_{i=1}^{m} Wx(yi)Hx(yi)$
       For 1≤i≤m
  4. Set OS = NULL
  5. For i = 1 to n do
  6. Obtain OS by calculating outlier factor OF(xo) using equation
     $OF(xo) = \sum_{i=1}^{m} OF(xo, i)$
  7. End for
  8. Build S by searching objects in OS
```

Above algorithm uses greedy approach for outlier detection.Algorithm firstly computesweighted entropy for each attribute [2]. Entropy of each attribute isupdated. Theattribute entropy is always changes when outliers aredetected and removed from the data set.And calculates outlier factor for each attribute and get thelargest OF set which will convert to OS (Outlier set).

After that set S will be built [2]. Complexity of the algorithm is O (nm), as we are not using any searching algorithm [2].

E. Experimental setup

The system is built using Java framework version jdk 1.7 on Windows platform. The Eclipse Indigo is used as a development tool. The system doesn't require any specific hardware to run, any standard machine is capable of running the application.

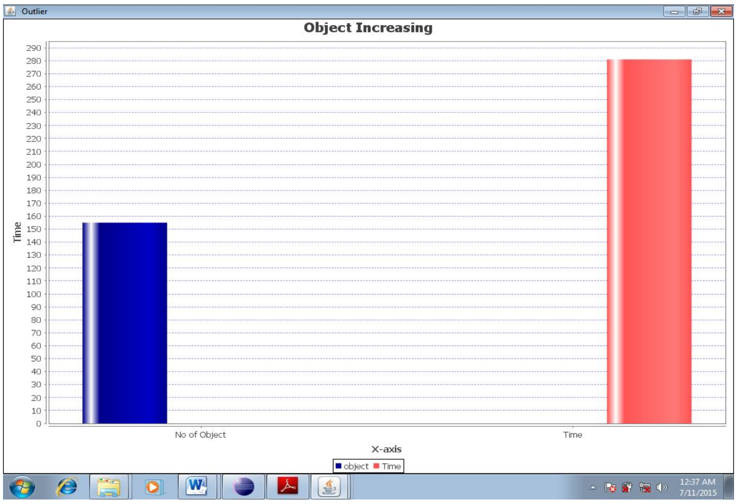
V. RESULTS AND DISCUSSION

Obj.	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	ITB-SP	Our algorithm
1	*k1*	*j2*	*k3*	*k4*	*e5*	*c6*	23.199	449.22
2	*j1*	*i2*	*j3*	*j4*	*b5*	*c6*	22.61	449.30
3	*i1*	*h2*	*i3*	*i4*	*d5*	*d6*	23.61	449.16
4	*h1*	*g2*	*h3*	*h4*	*c5*	*b6*	21.77	449.42
5	*g1*	*f2*	*g3*	*g4*	*b5*	*a6*	20.22	449.63
6	*f1*	*e2*	*f3*	*f4*	*a5*	*b6*	22.56	449.30
7	*a1*	*a2*	*c3*	*c4*	*a5*	*a6*	22.78	449.27
8	*a1*	*b2*	*a3*	*a4*	*f5*	*e6*	22.43	449.32
9	*a1*	*a2*	*a3*	*b4*	*g5*	*f6*	22.07	449.36
10	*c1*	*b2*	*b3*	*d4*	*h5*	*g6*	22.09	449.36
11	*d1*	*d2*	*b3*	*a4*	*i5*	*h6*	22.65	449.28
12	*b1*	*c2*	*d3*	*e4*	*g5*	*i6*	23.39	449.18
13	*e1*	*c2*	*e3*	*b4*	*k5*	*g6*	22.90	449.25

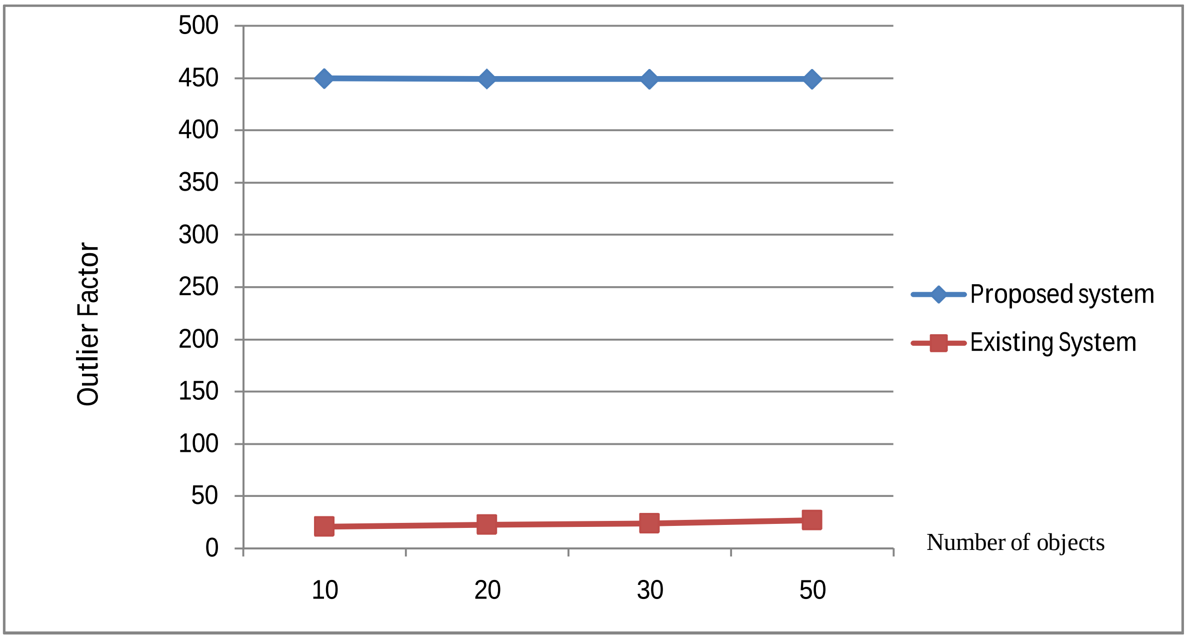
Table 5.1 Outlier factors of different systems on categorical data

Outlier factors of different methods are compared to gain a better understanding of the advantage of the proposed methods. The 13 objects are different from each other. We are computing OF (Outlier factor) for given categorical data set. Previous existing methods like ITB-SP are used to compute outlier factor for above data set and compare values of OF with proposed system. As shown in above Table 9.1, OF for attributes are varied from method to method. Specifically, the column ITB-SP shows different attributes with different OF values. Our system provides most precise assessment. It indicates that object 2 in the data set is less likely to be an outlier than objects 5, 6, 8, and 9, which are similar to each other. Moreover, objects 4 and 13 are likely to be outliers than objects 5-12, each of which is similar to only two other objects. These differences are important indices used by our system to accurately identify the most likely outlier candidate. We can also deal with real data sets like Zoo, Labor data efficiently using our algorithm.

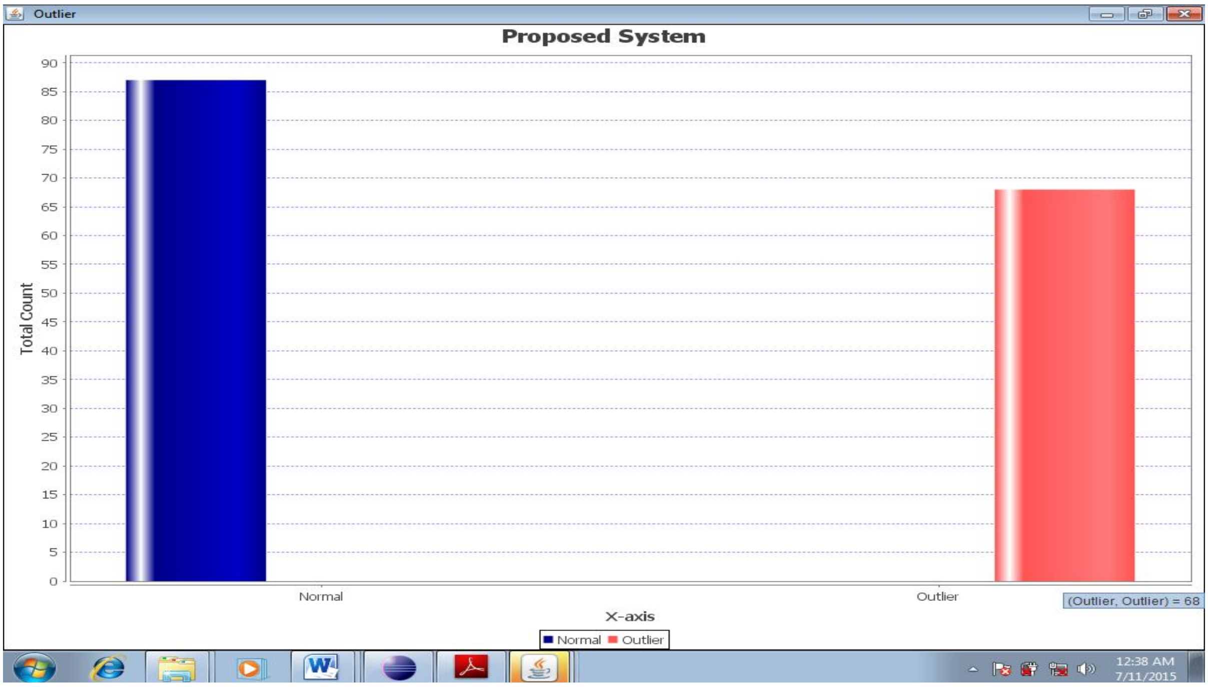
1. No. of objects and time required



2. *Comparison Graph*



3. *No. of outliers and normal objects from given file*



DOI: 10.15680/ijircce.2015.0306165

VI. CONCLUSION

This paper discusses many outlier detection methods based on information theory. We are proposing novel method which will overcome drawbacks of previous approaches. This paper formulates outlier detection as an optimization problem and proposed a practical, unsupervised, parameter less algorithm for detecting outliers in large-scale categorical data sets. Effectiveness of our approach results from a new concept of surfeit entropy. The efficiency of our algorithms results from the outlier factor function derived from the differential entropy. In particular, in our proposed approach dual total correlation and surfeit entropy works more effectively to remove outlier from large scale categorical attributes.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” ACM Computing Surveys, vol. 41, no. 3, pp. 1-58, 2009.
[2] V.J. Hodge and J. Austin, “A Survey of Outlier Detection Methodologies,” Artificial Intelligence Rev., vol. 22, no. 2, pp. 85- 126, 2004.
[3] E.M. Knorr and R.T. Ng, “Algorithms for Mining Distance-Based Outliers in Large Data Sets,” Proc. 24rd Int’l Conf. Very Large Data Bases (VLDB ’98), 1998.
[4] S.R. Gaddam, V.V. Phoha, and K.S. Balagani, “K-Means+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-Means Clustering and ID3 Decision Tree Learning Methods,”IEEE Trans. Knowledge and Data Eng., vol. 19, no. 3, pp. 345-354, Mar. 2007.
[5] V. Chandola, Banerjee, and Kumar, “Anomaly Detection: A Survey ”, ACM Computing Surveys, vol. 41, no. 3, pp. 1-58, 2009.
[6] V. Hodge and J. Austin “A Survey of Outlier Detection Methodologies,” Artificial Intelligence Rev, vol. 22, no. 2, pp. 85-126, 2012.
[7] J. Zhang,” Advancements of Outlier Detection: A Survey,”ICST Transactions on Scalable Information Systems, Vol.3,no. 1, March 2013.
[8] Sheu Wu, Member IEEE, and ShengruiWang,Member IEEE,“Information-Theoretic Outlier Detection for Large-ScaleCategorical Data,” IEEE transactions on knowledge and dataengineering, vol. 15, no. 3, march 2013.
[9] T. Cover and J. Thomas, “Elements of Information Theory”John Wiley & Sons, 1991, pp.12-21.
[10] AymanTaha, AliHadi,” A General Approach forAutomating Outliers Identification in Categorical Data, IEEEconference on Computer System & applications, pp.1 – 8,May 2013.
[11] S. Li, R. Lee, “Mining Distance-Based Outliersfrom Categorical Data,” Proc. IEEE Seventh InternationalConference, Data Mining Workshops (ICDM ’07), 2007.
[12] S.D. Bay “Mining Distance-BasedOutliers in Near Linear Time with Randomization and aSimple Pruning Rule,” Proc. Ninth ACM SIGKDDInternational Conf. Knowledge Discovery and Data Mining(KDD ’03), 2008.
[13] Z. He, X. Xu, Z.J. Huang, “FP-Outlier:Frequent Pattern Based Outlier Detection,” ComputerScience and Information Systems, vol. 2, pp. 103-118, 2005.
[14] M.E. Otey, A. Ghoting, “FastDistributed Outlier Detection in Mixed-Attribute Data Sets,”Data Mining and Knowledge Discovery, vol. 12, pp. 203-228, 2006.
[15]] H.D.K. Moonesignhe,Tan, “Outlier Detection UsingRandom Walks,” Proc. IEEE 18th Int’l Conf. Tools withArtificial Intelligence (ICTAI ’09), 2009.
[16] W. Qian, H. Lu, and A. Zhou, “Finding CentricLocal Outliers in Categorical/Numerical Spaces, Knowledgeand Information Systems, vol. 8, no. 3, pp 309-338, 2006.
[17] Nicholas Timme,_ Wesley Alford, Benjamin Flecker, andJohn M. Beggs,” Multivariate information measures: anexperimentalist’s perspective”, 28 Nov 2011.
[18] S. Watanabe, “Information Theoretical Analysis ofMultivariate Correlation,” IBM J. Research andDevelopment, vol. 4, pp. 66-82, 1960.