http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 11 (7): 782-787, 2012
ISSN 1812-5638 / DOI: 10.3923/itj.2012.782.787

Anomaly Detection in Transactional Sequential Data

[1]Jingwei Zhang, [1]Yuming Lin, [1]Huibing Zhang and [2]Qing Yang
[1]School of Computer Science and Engineering, Guilin University of Electronic Technology, Guilin, 541004, China
[2]Electronic Engineering and Automation Institute, Guilin University of Electronic Technology, Guilin, 541004, China

Abstract: A transactional system often has a steady distribution based on their decisions or responses on various transactions. If the system is changed, the distribution also often changes. It is a valuable work to detect anomaly caused by system changes based on different distributions. In this study, we modeled those decision or response signals into a series of time-related distributions and then proposed a method combining distance metrics and anomaly detection to discover whether changes have happened in some systems. Distance metrics on different distributions can decide whether changes have happened and anomaly detection can find what happened further. Extensive experiments show that our method has a good performance and can locate the anomaly accurately.

Key words: Distribution distance, indicator distribution, distance metrics

INTRODUCTION

For online transactions, whether a transaction will success are influenced by many factors. For example, you are using your cell phone for dialing, whether the transaction could continue depends on whether you have input a correct phone number, whether your account has enough money, whether the network is good, whether the receiver is in a region with signals and so on. Though a failure transaction maybe caused by systems themselves or other factors, a steady transactional system should have a steady distribution on its responses from a statistical perspective. In fact, the online transaction systems are always improved so that they can provide richer and more reliable services for their customers; it is also possible that some mistakes are left in those systems because of some casual actions. Those inadvertent mistakes are very difficult to discover since they only happen in some specific conditions and are not visible for all customers. For example, you maybe often get such message, You dialed cannot be connected, when dialing by your cell phone, obviously, this message cannot give us any help about how to do in the next step. Though those uncertain messages are inevitable, they should have a steady distribution in a large number of historical records. If some operators have a high probability on those uncertain messages suddenly, maybe the operators changed their systems.

It is an important problem for the third parties to monitor whether changes are made in the systems of any party involved in transactions. Anomaly detection on distributions is an effective mechanism for monitoring. In this study, we consider those response signals as indicators; all indicators should have a steady distribution if the corresponding systems are not changed for a large scale of transactions. A successful change on systems should let the systems provide better services and present a steady distribution, a bad change will cause the systems show unsteady indicator distributions. The problem of detecting system changes and system stability is transformed into computation of different indicator distributions.

Classification and clustering are two primary technologies for data analysis. Hannan *et al*. (2008) applied classification method on tire pressure and temperature data for intelligent vehicle performance evaluation. Hussain *et al*. (2008) provided a method to detect drowsiness through combining the Electrooculogram and Head Nodding signals, which can be analyzed and represented as feature vector of two states, alertness and sleepiness. Arora *et al*. (2009), Ranjan and Khalil (2007), Premalatha and Natarajan (2010) applied clustering methods on different data. Anomaly detection, which is similar to outlier detection, skyline query in some applications, such as network or intrusion detection (Thottan and Ji, 2003; Teodoro *et al*., 2009),

Corresponding Author: Qing Yang, Electronic Engineering and Automation Institute, Guilin University of Electronic Technology, No. 1, Jinji Road, Qixing District, Guilin, 541004, China Tel: 86-773-2291522

privacy preserving (Vaidya and Clifton, 2004), Software bugs (Hangal and Lam, 2002) and so on. Elahi *et al*. (2009) researched how to detect outstanding outliers from datastreams by grid-based technology. Abe *et al*. (2006) proposed a selective sampling mechanism based on active learning for outlier detection, Stibor *et al*. (2005) and Gnozalez *et al*. (2002) used classification methods for anomaly detection and focused on negative selection algorithm. Song *et al*. (2007) made use of users' domain knowledge to discover anomalies in which users are really interesting, this is a common goal with our paper. Chan and Mahoney (2005) provided some algorithms for anomaly detection on monitoring applications, in which time factor is considered. More information on anomaly detection is covered by Chandola *et al*. (2009). Uysal (2007) presented some applications and comparison of two models on the prediction of time series values, Radial Basis Function Networks and Autoregressive Integrated Moving Average models. Sharma *et al*. (2007) researched the interestingness measure to extract the interesting results from large number of classification results.

Present contribution is to model system stability as indicator distributions and use distance metrics and anomaly detection on different distributions to discover the changes of systems. Based on a large number of the indicators, our method can judge whether changes happened on corresponding systems.

PROBLEM SETTING

From statistically significant, the behaviors of a system can be represented as a series of distributions on their indicators. Given an indicator set $I = \{i_1, i_2, ..., i_n\}$ and two distributions on I, D_1 and D_2, $D_1 = \{<i_1, p_{11}>, <i_2, p_{12}>, ..., <i_n, p_{1n}>\}$, $D_2 = \{<i_1, p_{21}>, <i_2, p_{22}>, ..., <i_n, p_{2n}>\}$:

$$\forall i, p_{1i} \geq 0, p_{2i} \geq 0, \sum_{i=1}^{n} p_{1i} = 1, \sum_{i=1}^{n} p_{2i} = 1$$

D_1 and D_2 correspond to different system states and reflect different system characteristics. Generally, for the given indicator set, only one of them represents a successful indicator, such as i_1 and other failure. Given a time set $T = \{t_1, t_2, ..., t_m\}$, if all distributions are time-related, we have $TD = \{<t_1, D_1>, <t_2, D_2>, ..., <t_m, D_m>\}$, the aim of anomaly detection is to find abnormal distribution and locate the abnormal indicators based on a series of given distributions representing the stability of a system, which can be considered as strong evidence of bad system changes.

Problem definition: Given a set of time-related distributions $TDS = \{TD_1, TD_2, ..., TD_m\}$, anomaly detection is to find whether there is a distribution change on TDS at the time series $\{T_1, T_2, ..., T_m\}$. A distribution change on time series T means that there are two distribution sets DS_1 and DS_2, the time set related with DS_1 is T_1 and DS_2 corresponds to T_2, $T = T_1 \cup T_2 \wedge T_1 \cap T_2 = \varnothing$ and they satisfy $MAX(T_1) < MIN(T_2)$ or $MAX(T_2) < MIN(T_1)$.

A distribution change can be also understood as a distribution transfer from a steady state to another steady state. If the second steady state is formed by some bad system changes, it is valuable to capture such state since the system should be corrected immediately. In fact, anomaly detection includes two stages, the first stage is to define some steady distributions on $TD_1, TD_2, ..., TD_m$, the second is to discover detailed anomaly through comparison between the steady distribution and current distribution TD_{m+1}. In the following part of this study, we will pay attention to those failure indicators since the successful indicator and the set of failure indicators are related.

Example 1: If a system has 17 indicators, which are expressed as $A_0, A_1, A_2, ..., A_{16}$. A_0 is the successful indicator and $A_1, A_2, ..., A_{16}$ are failure indicators. In Fig. 1a, there are 5 different indicator distributions, their related time is $t_1<t_2<t_3<t_4<t_5$, indicators A_1, A_4 and A_7 are becoming high, which shows that the system may be changed. Figure 1b presents the comparison between baseline, which is established by historical distributions and current distribution TD_5, A_1, A_4 and A_7 are detected as anomaly since their distributions are higher than baseline's.

BASIC AND IMPROVED METHOD

Distance metrics is used to judge whether a current indicator distribution is not consistent with the baseline distribution and anomaly detection is used to discover which indicator is bad.

There are two situations for anomaly detection, one is that the successful indicator has a great change and the second is the successful indicator has no apparent change. Usually, only the first situation, namely the good indicator has a sharp decline, will be concerned. In this study, we will also introduce distance metrics to detect the second situation.

Distance metrics: Distance metrics are often used to cluster the similar points in multi-dimensional space. The classical distance metrics include Dice distance, Jaccard distance, Cosine distance, Euclidean distance and so on. These distance metrics have a wide range of applications, such as clustering analysis (Schenker *et al*.,

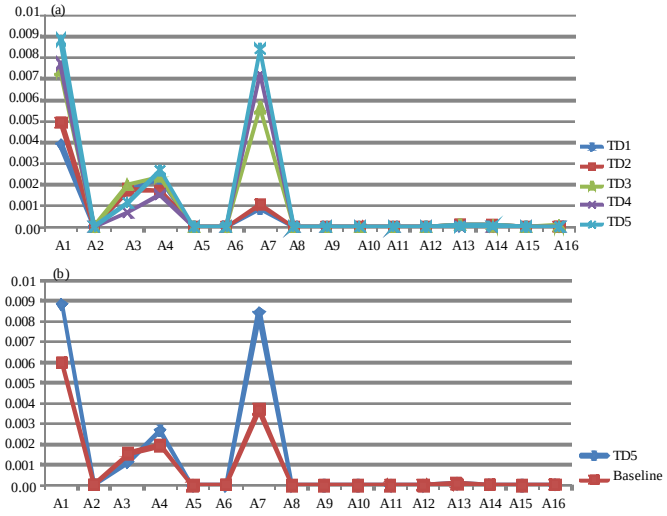

Fig. 1(a-b): An illustration of indicator distribution and anomaly detection, (a): Time-related indicator distributions, (b): A comparison of indicator distributions

2003), natural language processing (Lee *et al*., 2005; Dumais *et al*., 1998), text retrieval (Kim and Choi, 1999), image recognition (Salleh *et al*., 2011) and so on. Here we consider a distribution on a set of indicators as a point of multi-dimensional space, because all dimensions are in a same range, which are all probability, a value between 0 and 1, Euclidean distance is the best to measure their similarity. Given a set of such points $P = \{p_1, p_2, ..., p_n\}$, $\forall_i$, p_i is a distribution on a set of indicators, we assume that all points in P are normal, which means that these points are gathered when system is steady and has no anomaly, if a point p has a far distance from all points in P, we can say that there is anomaly in p.

In fact, it is difficult to find enough good distributions for anomaly detection, especially for the third parties since they cannot know whether a change has been made on systems. The usual situation is that we have a group of distributions, we need to distinguish those abnormal distributions from normal ones. In order to find which point is outlier, we use Euclidean distance to compute the distance between points and then group those points with small distance into a set, obviously, every set of points should belong to a steady distribution. Time information is checked to judge whether a system has a change. If we get two sets, S_1 and S_2, after distance computation, $S_1 = \{<t_{11}, p_{11}>, <t_{12}, p_{12}>, ..., \{t_{1n}, p_{1n}\}\}$, $S_2 = \{<t_{21}, p_{21}>, <t_{22}, p_{22}>, ..., \{t_{2m}, p_{2m}\}\}$, the time information related with them are $T_1 = \{t_{11}, t_{12}, ..., t_{1n}\}$ and $T_2 = \{t_{21}, t_{22}, ..., t_{2m}\}$, if MAX ($T_1$) < MIN ($T_2$) or MAX ($T_2$) < MIN ($T_1$), we can say there is a change in corresponding system. Here, we only consider MAX (T_1) < MIN (T_2), if all failure indicators in S_2 have lower probability than S_1, we say that this is a successful change on corresponding system, our aim is to find those failed change on systems, which means that some indicators in S_2 have a higher probability than S_1.

The partition is done by distance computation, we firstly pick up the point pair with the maximum distance, the two points are considered as the first element of corresponding categories. All remaining points are judged their owners on their average distance with current categories, a point will be allocated to the set with minimum average distance. The average distance between a point p and a set S is defined as:

$$Dis(p,S) = \frac{\sum \sqrt{\sum_i (p_j - S_{i,j})^2}}{|S|}$$

The detailed process for distribution partition is presented in algorithm 1.

Example 2: Table 1 illustrates five distributions on an indicator set $\{A_0, A_1, A_2, A_3\}$, which are $\{<t_1, D_1>, <t_2, D_2>, <t_3, D_3>, <t_4, D_4>, <t_5, D_5,>\}$ t_1 to t_5 is an increasing time sequence.

According to algorithm 1, five distributions can be divided into $S_1 = \{<t_1, D_1>, <t_2, D_2>, <t_3, D_3>\}$ and $S_2 = \{<t_4, D_4>, <t_5, D_5>\}$. If A_0 represents the successful indicator, there is a change on corresponding system since S_2 presents different distribution from S_1.

Algorithm 1: Abnormal distribution detection

Input: A set of distribution P = {p_1, p_2, ...,p_n}
Output: Distribution partition, C_1 and C_2

1: Computing distance between any two points based on Enclidean distance
2: Get the two points, p_1 and p_2, with maximum distance and establish two sets, C_1 ={p_1}and C_2={p_2}, remove p_1, p_2 from P
3: WHILE (|P|>0)
4: Get the next point p and remove p from P
5: IF Dis (p, C_1) <Dis (p, C_2)
6: C_1 = C_1∪{p}
7: Else
8: C_2 = C_1∪{p}
9: End if
10: End while
11: RETURN C_1, C_2

Table 1: Distributions on an indicator set

Indicators	A_0	A_1	A_2	A_3
t_1	0.99345	0.01020	0.0040	0.00235
t_2	0.98027	0.00865	0.0043	0.00678
t_3	0.98540	0.00900	0.0039	0.00170
t_4	0.96320	0.02100	0.0127	0.00310
t_5	0.95866	0.01820	0.0153	0.00784

Anomaly detection: According to distance metrics, we can only identify whether there are some changes on systems but we don't know the change details, namely which indicator has an offset on a series of distributions. In this section, we will discover the change details through anomaly detection. In fact, distance metrics measure changes from multi-dimensional perspective and anomaly detection find change details based on concrete one-dimensional perspective.

We use distribution trends to detect anomaly before and after system change. Whether an indicator is abnormal depends on whether it has different distribution trends before and after system changes. Here, distribution trends are measured by baseline and fluctuation range. Baseline is represented by average and fluctuation range is represented by variance. Given two distributions of a same indicator, which correspond to two phases, before and after system changes, we can compute their average, $base_1$ and $base_2$ and variance, var_1 and var_2, if the average has an obvious increase or variance have different change, we can say that the indicator is abnormal. In fact, an abnormal indicator except successful indicator has two forms:

- $base_2 > base_1 + \beta$, namely baseline has increased beyond the expected range after system change
- $\frac{var2}{var1} > \varepsilon$, indicator presents an unsteady state

For the successful indicator, the first rule should be $base_2 < base_1 + \beta$ and the second rule is right. Here, β and ε are threshold which are set experimentally. The detailed process is presented in algorithm 2.

Algorithm 2: Anomaly detection

Input: Two distributions P_1 and P_2 on an indicator Set S
Output: An abnormal indicator set, AS

1: For every indicator i∈S
2: Compute the average and variance on P_1 and P_2, remembered as p_{i1}, p_{i2} and var_{i1}, var_{i2}
3: If $\frac{var_{i2}}{var_{i1}} > \varepsilon$
4: AS = AS∪ {i}
5: Endif
6: If i is not successful indicator
7: If $p_{i2} > p_{i1} + \beta$
8: AS = AS∪ {i}
9: Endif
10: Else if $p_{i2} < p_{i1} + \beta$
11: AS = AS∪ {i}
12: End if
13: End for
14: Retum AS

Improvement on distance metrics and complexity: Though a steady system often presents a steady distribution on its indicators, exception is inevitable, which means that a distribution seems to be abnormal but no system change, strict distance metrics are not very accurate to solve this problem.

Since, it is possible that a distribution before change is divided into the distribution set after change and vice versa. Here, we introduce probability threshold to solve this question, suppose time sets T_1 and T_2 are collected from two time-related distribution sets S_1 and S_2, S_1 corresponds to a distribution before system change and S_2 is a distribution after system change, if:

$$\frac{\left(\left|\{t_i \mid \forall i, t_i \in T_2 \wedge t_i < MAX(T_1)\}\right| + \left|\forall i, t_i \in T_1 \wedge t_i > MIN(T_2)\}\right|\right)}{|T_1| + |T_2|} < \lambda$$

we still say that there is a change on corresponding system. Here λ is set empirically.

Another improvement is on computation complexity. Strict distance metrics need distance computation between any pair of distributions, whose time complexity is O (n^2). Here, since it is most possible to have the maximum distance between two distributions that are at the starting and end of a time series, we can pick up these two points for computation to establish the initial categories. Especially, the pair distance can be pre-computed and stored in a hash structure, which will improve the partition speed.

EXPERIMENTS

Here, we carried out experiments to verify the efficiency and effectiveness of the proposed method. The experiments use two datasets, one is transformed from real

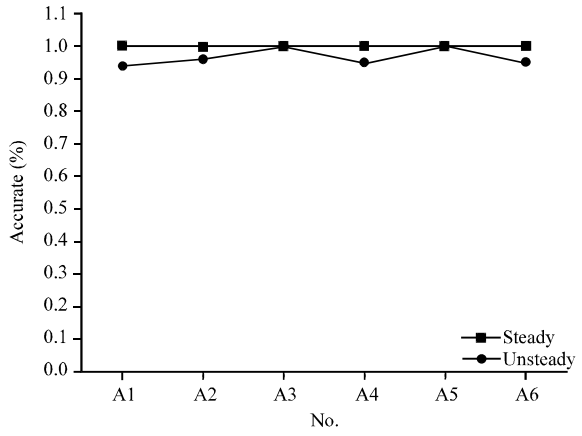

Fig. 2: Accurateness under steady and unsteady distributions

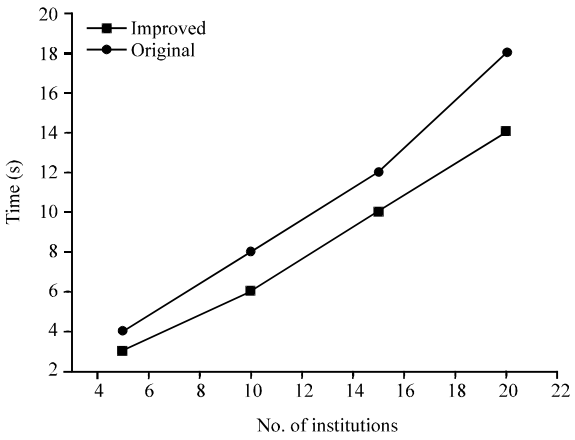

Fig. 3: Computation time comparison between original and improved program

transactional data and the other is generated by machine based on the former. All experiments run under Core Duo 2.2 GHz CPU and 2 G memory.

Accurateness: In order to verify the effectiveness of our method, we carried out a group of experiments on the first dataset, which is a transformed dataset on a real one. This dataset involves 6 institutions and 30 days' transactions. The number of transactions in one day varies from tens of thousands to several millions and the numbers of indicators for them are from 30 to 40. Here, we assume that every institution corresponds to a system since we cannot know how many systems really exist in an institution. Every experiment is carried out in two phases, this first phase is to find whether there is change on all institutions and the second phase is to find the change details. We compare the accurateness on two situations, one is to test whether the proposed method will find anomaly under steady distribution, another is to test whether the proposed method can locate the anomaly accurately under some distribution after system changes. The experimental results are presented in Fig. 2. We can find that when all distributions are steady, namely no system change, the proposed method is perfect since it does not give any false positives. When checking whether there is anomaly under system change, the proposed method maintains a high accurateness above 90%.

Time performance: In this group of experiments, we expand the dataset to verify the efficiency of our method. An expanded dataset is used in this group of experiments, we expanded the number of institutions to 5000, 10000, 15000 and 20000, every of which still has 30 days' transactions. The number of indicators still varies from 30 to 40. In fact, the original program has a good performance but after improvement, the performance presents a better effect. The experimental results are shown in Fig. 3. The original program has a quicker rise on execution time than the improved when the number of institutions becomes larger. It is possible to consume more time when the number of transaction day increases greatly. In this dataset, we only expanded the number of institutions base on their distributions.

CONCLUSIONS

In this study, we consider a steady system as a steady indicator distribution, so a system change can be detected through distribution offset. All information returned by a system are modeled into an indicator distributions, through monitoring the indicator distribution, a system change can be discovered effectively. We use a two-stage method to locate system change, the first stage helps to decide whether there is a change on system, the second stage detect which indicator has an abnormal distribution. The proposed method can help to find system change effectively, which is very significant for quality analysis of online transactions.

ACKNOWLEDGMENT

We gratefully acknowledge the support of Education Department Foundation of Guangxi under grants No. 201010LX154.

REFERENCES

Abe, N., B. Zadrozny and J. Langford, 2006. Outlier detection by active learning. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20-23,2006, ACM Press, USA., pp: 504-509.

Arora, A., S. Upadhyaya and R. Jain, 2009. Integrated approach of reduct and clustering for mining patterns from clusters. Inform. Technol. J., 8: 173-180.

Chan, P.K. and M.V. Mahoney, 2005. Modeling multiple time series for anomaly detection. Proceedings of the 5th IEEE International Conference on Data Mining, November 27-30, 2005, IEEE Computer Society, Washington, DC, USA., pp: 90-97.

Chandola, V., A. Banerjee and V. Kumar, 2009. Anomaly detection: A survey. ACM Computing Survey, 41: 58-58.

Dumais, S.T., J. Platt, D. Heckerman and M. Sahami, 1998. Inductive learning algorithms and representations for text categorization. Proceedings of 7th ACM International Conference on Information and Knowledge Management, November 2-07, 1998, Bethesda, MD, pp: 148-155.

Elahi, M., L. Xinjie, M.W. Nisar and H. Wang, 2009. Distance based outlier for data streams using grid structure. Inform. Technol. J., 8: 128-137.

Gnozalez, F., D. Dasgupta and R. Kozma, 2002. Combing negative selection and classification techniques for anomaly detection. Proceedings of the Congress on Evolutionary Computation (CEC`02), May 12-17, 2002, IEEE Computer Society, Honolulu, USA.,-pp: 705.

Hangal, S. and M. S. Lam, 2002. Tracking down software bugs using automatic anomaly detection. Proceedings of the 24th International Conference on Software Engineering, May 19-25, 2002, ACM, New York, pp: 291-301.

Hannan, M.A. A. Hussain, A. Mohamed and S.A. Samad, 2008. TPMS data analysis for enhancing intelligent vehicle performance. J. Applied Sci., 8: 1926-1931.

Hussain, A., B. Bais, S.A. Samad and S.F. Hendi, 2008. Novel data fusion approach for drowsiness detection. Inform. Technol. J., 7: 48-55.

Kim, M.C. and K.S. Choi, 1999. A comparison of collocation-based similarity measures in query expansion. Inf. Process. Manage., 35: 19-30.

Lee, C., H. Park and C. Ock, 2005. Significant sentence extraction by euclidean distance based on singular value decomposition. Proceedings of the 2nd International Joint Conference of Natural Language Processing, October 11-13, 2005, Springer-Verlag, Berlin Germany, pp: 636-645.

Premalatha, K. and A.M. Natarajan, 2010. A literature review on document clustering. Inform. Technol. J., 9: 993-1002.

Ranjan, J. and S. Khalil, 2007. Clustering methods for statistical analysis of genome databases. Inform. Technol. J., 6: 1217-1223.

Salleh, S.S., N.A.A. Aziz, D. Mohamad and M. Omar, 2011. Combining Mahalanobis and Jaccard to improve shape similarity measurement in sketch recognition. Proceedings of the 13th Internametion Conference on Computer Modelling and Simulation, March 30-April 1, 2011, IEEE Computer Society, Washington, DC, USA., pp: 319-324.

Schenker, A., M. Last, H. Bunke and A. Kandel, 2003. Comparison of distance measures for graph-based clustering of documents. Proceedings of the 4th IAPR International Conference on Graph Based Representations in Pattern Recognition, June 30-July 02, 2003, Springer-Verlag, Berlin, Germany, pp: 202-213.

Sharma, S., S. Khare and S. Sharma, 2007. Measuring the interestingness of classification rules. Asian J. Inform. Manage., 1: 43-49.

Song X., M. Wu, C. Jermaine and S. Ranka, 2007. Conditional anomaly detection. IEEE Trans. Knowl. Data Eng., 19: 631-645.

Stibor, T., P. Mohr and J. Timmis, 2005. Is Negative Selection Appropriate for Anomaly Detection. ACM, New York, pp: 504-509.

Teodoro, P.G., J.D. Verdejo, G.M. Fernandez and E. Vazquez, 2009. Anomaly-based network intrusion detection: Techniques, systems and challenges. Comput. Secur., 28: 18-28.

Thottan, M. and C. Ji, 2003. Anomaly detection in IP networks. IEEE Trans. Signal Proc., 51: 2191-2204.

Uysal, M., 2007. Comparison of ARIMA and RBFN models to predict the bank transactions. Inform. Technol. J., 6: 475-477.

Vaidya, J. and C. Clifton, 2004. Privacy-preserving outlier detection. Proceedings of the 4th IEEE International Conference on Data Mining (ICDM 2004), November 1-4, 2004, Brighton, UK., pp: 233-240.