Research Journal of Applied Sciences, Engineering and Technology 6(8): 1340-1347, 2013
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: August 02, 2012 Accepted: September 03, 2012 Published: July 10, 2013

Outlier Detection Scoring Measurements Based on Frequent Pattern Technique

Aiman Moyaid Said, Dhanapal Durai Dominic and Brahim Belhaouari Samir
Department of Computer and Information Sciences, Faculty of Science and Information Technology, Universiti Teknologi PETRONAS, Tronoh, Perak, Malaysia

Abstract: Outlier detection is one of the main data mining tasks. The outliers in data are more significant and interesting than common ones in a wide variety of application domains, such as fraud detection, intrusion detection, ecosystem disturbances and many others. Recently, a new trend for detecting the outlier by discovering frequent patterns (or frequent item sets) from the data set has been studied. In this study, we present a summarization and comparative study of the available outlier detection scoring measurements which are based on the frequent patterns discovery. The comparisons of the outlier detection scoring measurements are based on the detection effectiveness. The results of the comparison prove that this approach of outlier detection is a promising approach to be utilized in different domain applications.

Keywords: Anomaly, frequent pattern mining, outlier detection, outlier measurement

INTRODUCTION

An outlier in a dataset is an observation or a data object that deviated so much from other observations as to arouse suspicion that it was generated by different mechanism or causes (Hawkins, 1980). Detecting outlier in the data is an important data mining task and required in many real applications, such as fraud detection, marketing analysis, weather predication and network intrusion. Commonly, the detection of the outlier procedure is divided into two steps, first is to define what the outlier should be in a given dataset and then find a method to detect these outliers.

Many outlier detection techniques have been developed and evaluated in the last several years (Gogoi *et al*., 2011; Chandola *et al*., 2009). In general the existing techniques for detecting outliers are suffering from the following drawbacks:

- High dimensional space curse
- High computational cost
- Given a reason for outliers

Recently, a new trend for detecting the outlier by discovering frequent patterns (or frequent item sets) from the data set has been studied. In this study, we present a summarization and comparison of the available outlier detection measurements which are based on the frequent patterns discovery.

OUTLIER MEASUREMENTS

Detection outliers in a static data: The problem of discovering the frequent item sets in transactional and relational database was first studied by Agrawal and Srikant (1994). The problem can be defined formally as follows:

Let $I = \{i_1, i_2, i_3, \ldots, i_n\}$ be a set of items and the database $D = \{t_1, t_2, \ldots, t_n\}$ be a set of transactions, each consisting of a set of item from I. An item set X is a non-empty subset of I. The length of the item set X is the number of items in X.

An item set containing i items is called an i-item set. A transaction $t \in D$ is said to contain item set X if $X \subseteq t$. The support of item set X is defined as:

$$\text{Support}(X) = \frac{\| \{t \in D \mid x \subseteq t\} \|}{\| \{t \in D\} \|} \in [0, 1]$$

The item set X holds in the transaction set D with support, which is the percentage of transactions that contain an item set. For an item set to be interesting, its support value must be higher than a user-specified minimum threshold. Such item set are said to be frequent item set (or frequent pattern).

Based on the frequent item sets discovery paradigm, several approaches which have been developed to find the outliers in the transactional database. These approaches are described in the following subsections:

Frequent Pattern Outlier Factor (FPOF) scoring measurement: Zengyou *et al*. (2005) has proposed a new method for detecting outliers by using frequent patterns. Frequent patterns represent the common patterns which occur in many data objects or to large

Corresponding Author: Aiman Moyaid Said, Department of Computer and Information Sciences, Faculty of Science and Information Technology, Tronoh, Perak, Malaysia, Tel.: 0060122591226

percentage of data objects. An extra number of frequent patterns in a data object imply that this object is probably normal data, because it possesses the "common features" of the dataset. In contrast, for a data object that contains less frequent patterns, it means that this data object is likely to be an outlier. Zengyou has presented a new measurement to be used as the basic metric for identifying outliers. In addition to identifying the outliers, another measurement was also presented to describe the reasons why identified outliers are abnormal. The measurement is defined as:

Let $D = \{t_1, t_2, ..., t_n\}$ be a database containing a set of n transactions with items I. Given a threshold minisupport, the set of all frequent patterns is denoted as: FPS (D, minisupport) = $\{X \subseteq I \mid \text{support}(X) \geq \text{minisupport}\}$. For each transaction t, the frequent pattern outlier factor of t is defined as:

$$\text{FPOF}(t) = \frac{\sum_{x \subseteq t,\, x \in \text{FPS}(D, \text{minisupport})} \text{support}(x)}{||\text{FPS}(D, \text{minisupport})||}$$

$\forall x \in t$ and $x \in$ FPS (D, minisupport), $||..||$ is the number of elements contained in the collection. The explanation of this measurement is, if a transaction t contains more frequent patterns, that mean more support value for this transaction and then its FPOF value will be big, which indicates that it is unlikely to be an outlier. In contrast, transactions with small FPOF values are likely to be outliers. The value of the FPOF is between 0 and 1. To describe the reasons why identified outliers are abnormal, the item sets not contained in the transaction (it is said that the item set is contradictive to the transaction) are good candidates for describing the reasons.

Contradict-ness measurement: For each transaction t, a frequent item set X is said to be contradictive to t if $X \not\subset t$. The contradict-ness of X to t is defined as:

$$\text{Contradict-ness}(X, t) = (||X|| - ||t \cap X||) * \text{support}(X)$$

The explanation of this measurement is, the first part (i.e., $(||X|| - ||t \cap X||)$) represents the number of element which is in X but not in t, means that longer item sets give better description than that of short ones. The second part (i.e., support (X)) represents the percentage frequency of this item set in the database, means the greater the support of the item set X, the greater the value of contradict-ness of X to t, since a large support value of X suggests a big deviation.

With measurement Contradict-ness it is possible to identify the contribution of each item set to the outlying-ness of the specified transaction, that is each item set in the FPS (D, minisupport) and not in that specified transaction. The list of all the contradict item set is enormous; therefore, it is not feasible to all list of them. A practical approach is to present only the top-k, k is integer number, contradict frequent patterns to the end user.

Weighted Closed Frequent Pattern Outlier Factor (WCFPOF) scoring measurement: In the study done by Ren *et al*. (2009), the researcher studied the drawbacks in the FindFPOF and proposed another algorithm which is considered as an improvement for FindFPOF algorithm. A drawback of the Find FPOF algorithm is that the size of the extracted frequent patterns is huge, because FindFPOF algorithm finds all the frequent patterns in the data set. Another drawback is, using the Apriori algorithm (Agrawal *et al*., 1993) as algorithm for finding the frequent patterns, which is time-consuming. Data object that contains more closed frequent patterns and the weights of the corresponding closed frequent patterns have comparatively great values. This means that this data object is more likely to be a normal data. In contrast, for a data object that contains less closed frequent patterns, this means that this data object is likely to be an outlier. The measurement is defined as:

Let $D = \{t_1, t_2, .., t_n\}$ be a database containing a set of n transactions with items I. Given a threshold minisupport, the set of closed frequent patterns is denoted as:

$$\text{CFPS}(D, \text{minisupport}) = \{X \subseteq I \mid \text{support}(X) \geq \text{minisupport} \wedge \nexists Y \supsetneq X \mid \text{support}(X) = \text{support}(Y)\}$$

For each transaction t, the weighted closed frequent pattern outlier factor of t is defined as:

$$\text{WCFPOF}(t) = \frac{\sum_{x \subseteq t,\, x \in \text{CFPS}(D, \text{minisupport})} \text{support}(x) \frac{|x|}{|t|}}{||\text{CFPS}(D, \text{minisupport})||}$$

where,

$\forall X \in t$ and $X \in$ CFPS (D, minisupport)

$||..||$: The number of elements contained in the collection

$|..|$: The length

$\frac{|x|}{|t|}$: The weight of item set X

If the closed frequent pattern is heavily influenced; the weight of the closed frequent pattern is close to 1.

The explanation of this measurement is, if a transaction t contains more closed frequent patterns and the weights of corresponding closed patterns have

comparatively great value, then its WCFPOF value will be big, which indicates that it is unlikely to be an outlier. In contrast, transactions with small WCFPOF values are likely to be outliers. The value of the WCFPOF is between 0 and 1.

Weighted contradict-ness measurement: For each transaction t, a closed frequent itemset X is said to be contradictive to t if $X \not\subset t$. The contradict-ness of X to t is defined as:

Contradict-ness (X, t) = (||X||-||t∩X||) *support (X) (|X/|t||)

The explanation of this measurement is, the first part (i.e., (||X||-||t∩X||)) stands for the number of element which are in X but not in t, means that longer item sets give better description than that of short ones. The second part (i.e., support (X)) represents the percentage frequency of this item set in the database, means the greater the support of the item set X, the greater the value of contradict-ness of X to t, since a large support value of X suggests a big deviation. The last part |X/|t|| is the weight of closed frequent item set X to t.

With measurement Contradict-ness it is possible to identify the contribution of each item set to the outlying-ness of the specified transaction, that is each item set which is in the CFPS (D, minisupport) and not in that specified transaction. Therefore, it is not feasible to list all the contradict item set, it is desirable to present only the top-k, k is integer number, contradictive closed frequent patterns to the end user.

This method targets two drawbacks of FindFPOF algorithm, it was intended to solve the problem of time-consumption, but implicitly it also solved the problem of redundant frequent patterns, which was studied by Zhang *et al*. (2010).

Longer Frequent Pattern Outlier Factor (LFPOF) scoring measurement: Another approach for overcoming the drawbacks of FPOF measurement was formulated by Zhang *et al*. (2010). In this approach, he suggested that data object that contains longer frequent patterns (i.e., longer superset) is more likely to be a normal data object, because it has more subset frequent patterns than other data objects. In contrast, a data object that contains short frequent patterns is more likely to be an outlier. The measurement is defined as:

Let $D = \{t_1, t_2, ..., t_n\}$ be a database contain a set of n transactions with items I. Given a threshold minisupport, the set of all frequent patterns is denoted as: FPS (D, minisupport) = {X⊆I |support (X) ≥minisupport}. Let F (t) be a frequent pattern set which composes of frequent patterns contained in t_i, where $t_i \in D$. F (t) is derived as $F(t) = \{X | X \in FPS \wedge t \supseteq X\}$. For each transaction t, the Longer Frequent pattern outlier factor of t is defined as:

$$LFPOF(t) = \frac{|X_{MAX}|}{|t|}$$

where, $|X_{MAX}|$ = The length of the longest frequent pattern in F(t) |t| = The length of the transaction t The explanation of this measurement is, if a transaction t contains a longer frequent patterns, then its LFPOF value will be big, which indicates that it is unlikely to be an outlier. In contrast, transactions with small LFPOF values are likely to be outliers. The value of the LFPOF is between 0 and 1.

Detection outliers in a dynamic data:

Frequent Pattern Outlier Factor (FPOF) scoring measurement: Zengyou *et al*. (2003) has employed the same idea from Zengyou *et al*. (2005) to find the outliers in the data stream environment. Therefore, data object that contains more frequent patterns, it means that this data object is more likely to be a normal data because it possesses the "common features" of the dataset. In contrast, a data object that contains less frequent patterns, this mean that this data object is likely to be an outlier.

Frequent Pattern Outlier Factor (FPOF) measurement and the contradict-ness measurement are used as basic measurement to identify outliers and to describe the reason why the identified outliers are abnormal.

The key aspect for detecting frequent patterns outliers is to get all the frequent item sets. However the existing methods for frequent pattern mining require multiple passes over the datasets, which is not allowed in the data stream model. Thus instead of finding the exact frequent patterns, the estimated frequent patterns is used by exploring approximation counts technique over data streams. The researcher utilized Lossy counting algorithm (Manku and Motwani, 2002) for finding the approximate frequent patterns from the data stream and then calculate the outlier factor (i.e., FPOF) for the data object using the frequent patterns seen so far. Due to the huge number of the frequent patterns, only the top k outliers are outputted to the user.

Weighted Frequent Pattern Outlier Factor (WFPOF) scoring measurement: Zhou *et al*. (2007) looked at the way to find the outliers in the data stream, which is totally different from the static data sets.

Another problem addressed by this research was the effect of the dimensionality on the accuracy of the outlier detection. A new outlier measurement was proposed. The idea of this measurement is the same as the FPOF measurement, the more frequent item sets the more normal the data object, otherwise the data object is outlier. Another drawback of FindFPOF algorithm is the speed of detecting the outliers in the data set. First for discovering the frequent patterns in the data set, at least two scans are needed and then after that, the data set needs to be scanned again to calculate the value of the FPOP measurement for each data object. Second for the approaches which are similar to FPOP measurement still targets the entire data set.

Therefore, Zhou presented a new algorithm which works under specific constraint of time and memory to be suitable for the data stream environment. This algorithm dynamically maintains the frequent patterns and also calculates the WFPOP value for each transaction, to determine whether it is outlier. This approach requires one scan for the data, which makes it efficient for data stream environment. The measurement is defined as:

The lengths of the frequent pattern which are belonging to the transaction are different; the longer frequent pattern is more meaningful. Therefore if you just consider the support of the frequent pattern obviously some useful information will be lost. To overcome this drawback (Zhou *et al*., 2007) improved the FPOF and defined the Weighted Frequent Pattern Outlier Factor (WFPOF) as:

$$\text{WFPOF}(t) = \frac{\sum_{x \subseteq t, x \in \text{FPS}(DS, minisupport)} \text{support}(x) \frac{|x|}{k}}{\| \text{FPS}(DS, minisupport) \|}$$

where,
|x| = The length of the pattern x
k = The dimensional of data space
FPS (DS, minisupport) = {x| support(x) ≥minisupport}

The explanation of this measurement is, if a transaction t contains a lot of long frequent patterns (i.e., heavy weight patterns) and a lot of frequent patterns, then its WFPOF value will be big, which indicates that it is unlikely to be an outlier. In contrast, transactions with small WFPOF values are likely to be outliers. Using the weight of the frequent pattern has a good effect on improving the accuracy of measuring the outlier degree for each data object. The value of the WFPOF is between 0 and 1.

Frequent Pattern Contradiction Outlier Factor (FPCOF) scoring measurement: Tang *et al*. (2009) studied the problem of detecting outliers in a sliding window of any size over online data streams. He proposed a new method for detecting outliers in online data stream by using frequent patterns.

If data object contains more frequent patterns with less contradictive to that data object over a sliding window; it could mean that this data object is more likely to be a normal data. In contrast, a data object that contains less frequent patterns with more contradictions to that data object, which means that this data object is likely to be an outlier. The measurement is defined as:

Let I = $\{i_1, i_2, ..., i_n\}$ be a set of all items, DS = $\{t_1, t_2, .., t_i, ..\}$ is a sequence of incoming transactions in the data stream where t_i is a transaction in the data stream, each consisting of a set of item from I. The sliding window, SW, contains the N latest transactions, where, N represents the width of the sliding window. An item set X in the sliding window SW whose width is N is said to be frequent if it has support greater than the minisupport. Given a threshold minisupport, the set of frequent patterns over the data stream is denoted as:

$$FPS\ (\text{SW, minisupport}) = \{Y \subseteq I \mid \text{support}(Y) \geq \text{minisupport} \wedge Y \in \text{SW}\}$$

For each transaction t, the frequent pattern contradiction outlier factor of t is defined as:

$$\text{FPCOF}(t) = \frac{\sum_{x \subseteq t, x \in \text{FPS}(SW, minisupport)} \frac{|x| - |x \cap t|}{|x|} \text{support}(x)}{\| \text{FPS}(SW, minisupport) \|}$$

where,
$X \in$ FPS (SW, minisupport)
||..||: The number of elements contained in the collection
|..| : The length

The explanation of this measurement is, the definition of the FPCOP shows the effect of the length of the frequent pattern. The more the value of the (|X|-|X∩ t|) / |X|, it means this frequent pattern is more contradictive to the transaction. Therefore it will result a big value for this transaction that contain a lot of these frequent patterns (i.e., contradictive frequent pattern). Consequently, the bigger the value of the FPCOF measurement the more likelihood it is considered as an outlier. In contrast, transactions with small FPCOF values are likely to be normal. The value of the FPCOP is between 0 and 1.

Maximal Frequent Pattern Outlier Factor (MFPOF) scoring measurement: Lin *et al*. (2010) in his research focused upon the problem to find the outliers in the high-dimensional time-series data stream. Another

problem was addressed by this research is the efficiency of discovering the frequent pattern from the data steam. If data object contains more frequent patterns, it could mean that this data object is more likely to be a normal data. In contrast, a data object that contains less frequent patterns could mean that this data object is likely to be an outlier. Any non-empty subset of maximal frequent pattern is a frequent pattern and it means that the maximal frequent item sets already implicitly contain frequent item sets. The number of the frequent pattern in the maximal frequent item sets is much less than that in frequent item sets. Thus if the data object has more intersection with maximal frequent patterns, it means that this data object is unlikely to be an outlier. The measurement is defined as:

$$\text{MFPOF}(t) = \frac{\sum_{x \subseteq t, x \in \text{MFPS}(D,minisupport)} \text{support}(x) \times |\, t \cap x \,|}{\| \text{MFPS}(D,minisupport) \|}$$

where, MFPS (D, minisupport) = The maximal frequent pattern of the dataset D. which is smaller than the complete set of frequent pattern. Therefore, this will help to reduce the computational complexity of this measurement.

The explanation of this measurement is, the definition of the MFPOF shows the effect of the number of maximal frequent patterns on the outlier factor. The more contains more maximal frequent patterns; it means that this data object is more likely to be a normal data. In contrast, a data object that contains less maximal frequent patterns, it means that this data object is likely to be an outlier. The value of the MFPOF is between 0 and 1.

ACCURACY COMPARISONS

In order to evaluate outlier detection scoring measurements summarized in this study, two accuracy measurements, the detection rate and detection precision are used (Narita and Kitagawa, 2008). Those measurements are defined as:

Detection rate = Number of detected true outliers/number of all true outliers

Detection precision = Number of detected true outliers/number of detected transactions as outliers

If outlier detection works well, it is expected that the rare classes would be over-represented in the set of points found. Three data sets are used to compare between the measurements (Intrusion dataset, breast-cancer-winsconsin dataset, lymphography dataset). Table 1 shows the dimensionality and the distribution of the three data sets.

We are comparing between five measurements approaches only, because the source code of WCFPOF measurement is not available. Table 2 and 3, show the detection rate and the precision results of the five measurements on Intrusion dataset respectively. As shown in those two tables, the best performing measurement is LFPOF. It can be seen from the data in the Table 2 and 3, MFPOF did not include any outlier in the top highest thirty scores, which gives it the worst effectiveness.

Table 4 and 5, show the detection rate and the precision results of the five measurements on Breast-cancer-winsconsin dataset. As displayed in those two tables the best performance is achieved by MFPOF. While for the performance of the WFPOF and FPOF measurements are similar and the performance of the LFPOF and FPCOF are almost comparable. The

Table 1: Class distribution for the three data sets

Data set	Dimensionality	Class	Number	(%)	Total
Intrusion	14	Normal	970	97	1000
		Outliers	30	3	
Breast-cancer-winsconsin	9	Normal	443	97	457
		Outliers	14	3	
Lymphography	18	Normal	142	96	148

Table 2: Detection rate result on the intrusion dataset

Top K	WFPOF	FPOF	LFPOF	FPCOF	MFPOF
5	0.10	0.10	0.10	0.10	0
10	0.17	0.17	0.23	0.17	0
15	0.33	0.33	0.40	0.33	0
20	0.50	0.50	0.57	0.50	0
25	0.67	0.67	0.70	0.67	0
30	0.80	0.80	0.87	0.80	0
Avg.	0.43	0.43	0.48	0.43	0

Table 3: Precision rate result on the intrusion dataset

Top K	WFPOF	FPOF	LFPOF	FPCOF	MFPOF
5	0.60	0.60	0.60	0.60	0
10	0.50	0.50	0.70	0.50	0
15	0.67	0.67	0.80	0.67	0
20	0.75	0.75	0.85	0.75	0
25	0.80	0.80	0.88	0.80	0
30	0.80	0.80	0.87	0.80	0
Avg.	0.69	0.69	0.78	0.69	0

Table 4: Detection rate result on the breast-cancer-winsconsin dataset

Top K	WFPOF	FPOF	LFPOF	FPCOF	MFPOF
4	0.22	0.22	0.22	0.29	0.29
6	0.29	0.29	0.36	0.43	0.43
8	0.43	0.43	0.50	0.58	0.58
10	0.58	0.58	0.58	0.58	0.70
12	0.60	0.60	0.60	0.60	0.86
14	0.60	0.60	0.79	0.60	0.90
Avg.	0.45	0.45	0.51	0.51	0.63

Table 5: Precision rate result on the breast-cancer-winsconsin dataset

Top K	WFPOF	FPOF	LFPOF	FPCOF	MFPOF
4	0.75	0.75	0.75	1	1
6	0.67	0.67	0.80	1	1
8	0.75	0.75	0.88	1	1
10	0.80	0.80	0.80	0.80	1
12	0.75	0.75	0.75	0.75	1
14	0.60	0.60	0.79	0.60	0.90
Avg.	0.71	0.71	0.80	0.83	0.98

Table 6: Detection rate result on the lymphography dataset

Top K	WFPOF	FPOF	LFPOF	FPCOF	MFPOF
2	0.30	0.30	0.30	0.30	0.30
3	0.50	0.50	0.50	0.50	0.50
4	0.67	0.67	0.67	0.67	0.67
5	0.67	0.67	0.80	0.67	0.80
6	0.67	0.67	0.80	0.67	0.80
7	0.80	0.80	0.80	0.80	0.80
Avg.	0.60	0.60	0.65	0.60	0.65

Table 7: Precision rate result on the lymphography dataset

Top K	WFPOF	FPOF	LFPOF	FPCOF	MFPOF
2	1	1	1	1	1
3	1	1	1	1	1
4	1	1	1	1	1
5	0.80	0.80	1	0.80	1
6	0.67	0.67	0.80	0.67	0.80
7	0.70	0.70	0.70	0.70	0.70
Avg.	0.86	0.86	0.92	0.86	0.92

number of detected true outliers found by MFPOF is similar to the number of detected transactions as outliers for the first twelve top outliers, which gives it the highest precision rate among the other scoring measurements.

Table 6 and 7, present the detection rate and the precision results of the five measurements on Lymphography dataset. As presented in those two tables, the performance of the WFPOF, FPOF, FPCOF measurements are similar and the performance of LFPOF and MFPOF is better than the others within top seven highest outliers (i.e., Top K).

From the comparisons, we conclude that the outlier detection scoring measurements which are based on the length of the pattern (LFPOF and MFPOF) outperform the measurements which are based only on the support of the entire frequent pattern.

DISCUSSION

Several studies have been done in the area of outlier detection using frequent pattern mining approach. This study set out with the aim of assessing the accuracy of those scoring measurements. As illustrated by the pervious section, the most interesting finding was that the outlier detection scoring measurements which are based on the length of the pattern outperform the other measurements. It can thus be suggested that this approach does not fit univariate data set efficiently. While it is valuable for finding outlier in multivariate data set with common patterns through all the data set. This aspect can address the problem of scoring measurements which are based on similarity measurement (i.e., Euclidean distance) with vast dimensionality. Most of the previous research adopted the number of the frequent patterns in the transaction to determine the score of the outlierness. While the results of the comparisons indicate that the measurements that are adapting the length the pattern are more accurate in detecting the outliers. All of the scoring measurements only adequate for binary attributes because they are based on different common patterns of attributes occur through the data set.

However, there are an open issues need to be addressed to enhance the performance of this approach, the first issue is to increase the accuracy of the detection and minimize the false positive rate taking into consideration the size and the dimensionality of the data. Another issue needs to be addressed; the need of measuring the degree of outlierness in the other type of attributes (i.e., numeric and categorical attributes). In addition to that, employing this approach in different application domains to verify its efficiency in a real application is substantial to be investigated.

CONCLUSION AND RECOMMENDATIONS

In contrast to traditional data mining task that aims to find the general pattern applicable to the majority of data, outlier detection targets the finding of the rare data whose behavior is very exceptional when compared with the rest large amount of data. In this study, we have attempted to give overview of the related study being done in outlier detection measurements which are based on frequent pattern mining. Several comparisons conducted on those measurements, the aim was to assess the detection accuracy. This study has found that generally those scoring measurements which are based on the length of the pattern have performed better than the rest. There are several measurements presented in this area and this approach of outlier detection is promising approach because it can deal with the high dimensional space of the data. Further research might explore these scoring measurements with different data type attributes (i.e., numerical and categorical) and focused research in this area promises useful applications.

REFERENCES

Agrawal, R. and R. Srikant, 1994. Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases, pp: 487-499.

Agrawal, R., T. Imieliński and A. Swami, 1993. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data. Washington, D.C., May 26-28.

Chandola, V., A. Banerjee and V. Kumar, 2009. Outlier detection: A survey. ACM Comput. Surv., pp: 1-72.

Gogoi, P., D.K. Bhattacharyya, B. Borah and J.K. Kalita, 2011. A survey of outlier detection methods in network anomaly identification. Comput. J., 54(4): 570-588.

Hawkins, D., 1980. Identification of Outliers. Chapman and Hall, Reading, London.

Lin, F., W. Le and J. Bo, 2010. Research on maximal frequent pattern outlier factor for online high dimensional time-series outlier detection. J. Convergence Inform. Technol., 5(10): 66-71.

Manku, G.S. and R. Motwani, 2002. Approximate frequency counts over data streams. Proceedings of the 28th International Conference on Very Large Data Bases, pp: 346-357.

Narita, K. and H. Kitagawa, 2008. Outlier detection for transaction databases using association rules. Proceeding of the 9th International Conference on Web-age Information Management (WAIM '08), pp: 373-380.

Ren, J., Q. Wu, C. Hu and K. Wang, 2009. An approach for analyzing infrequent software faults based on outlier detection. Proceeding of International Conference on Artificial Intelligence and Computational Intelligence. Shanghai, 4: 302-306.

Tang, X., G. Li and G. Chen, 2009. Fast detecting outliers over online data streams. Proceeding of International Conference on Information Engineering and Computer Science. Wuhan, pp: 1-4.

Zengyou, H., X. Xiaofei and D. Shengchun, 2003. Outlier detection over data streams. Proceeding of the 7th International Conference for Young Computer Scientists (ICYCS'03).

Zengyou, H., X. Xiaofei, J.Z. Huang and S. Deng, 2005. FP-outlier: Frequent pattern based outlier detection. Comput. Sci Inform. Syst., 2(1): 103-118.

Zhang, W., J. Wu and J. Yu, 2010. An improved method of outlier detection based on frequent pattern. Proceeding of WASE International Conference on Information Engineering. Beidaihe, Hebei, 2: 3-6.

Zhou, X.Y., Z.H. Sun, B.L. Zhang and Y.D. Yang, 2007. A fast outlier detection algorithm for high dimensional categorical data streams. J. Softw., 18(4): 933-942.