Associating Absent Frequent Itemsets with Infrequent Items to Identify Abnormal Transactions

Li-Jen Kao
Department of Computer Science and Information Engineering
Hwa Hsia Institute of Technology
New Taipei City, Taiwan 23568
lijenkao@cc.hwh.edu.tw

Yo-Ping Huang*
Department of Electrical Engineering
National Taipei University of Technology
Taipei, Taiwan 10608
yphuang@ntut.edu.tw
*corresponding author

Frode Eika Sandnes
Institute of Information Technology
Faculty of Technology, Art and Design
Oslo and Akershus University College of Applied Sciences
Oslo, Norway
Frode-Eika.Sandnes@hioa.no

Abstract Data stored in transactional databases are vulnerable to noise and outliers and are often discarded at the early stage of data mining. Abnormal transactions in the marketing transactional database are those transactions that should contain some items but do not. However, some abnormal transactions may provide valuable information in the knowledge mining process. The literature on how to efficiently identify abnormal transactions in the database as well as determine what causes the transactions to be abnormal is scarce. This paper proposes a framework to realize abnormal transactions as well as the items that induce the abnormal transactions. Results from one synthetic and two medical data sets are presented to compare with previous work to verify the effectiveness of the proposed framework.

Keywords data mining; abnormal transactions; absent frequent itemset; infrequent items; association rules.

1 Introduction

Data mining is an emerging technology used to discover interesting patterns from large databases. In the past more efforts reported in literature were focused on developing efficient methods to find association rules [5, 9-11, 14, 30, 32]. Recently, outlier detection attracts attention due to its importance in detecting deviant data. Several applications rely on outlier detection for the discovery of vital information such as credit card fraud detection, network intrusion detection, abnormal numeric values in stock prices, and disease symptom diagnosis [1-2, 6-7, 12, 17-18, 28, 30-31, 34-36].

Though some breakthroughs have been reported on outlier detection, there remain critical issues to be resolved. First, most outlier detection algorithms are designed for numerical data and rely on computing the relative distance between data points. These algorithms are not suitable for datasets with categorical attributes [17]. The following example illustrates why distance measuring methods fail to detect outliers in categorical datasets. Table 1 contains 10 transactions that can be divided into 3 types, i.e., {*item*1, *item*2}, {*item*1} and {*item*2}. If the minimum support and minimum confidence are set to 50% and 80%, respectively, the rule *item*1→ *item*2 will be an association rule, instead of the rule *item*2→ *item*1. According to the derived association rule, if someone buys *item*1, it is very possible that they will buy *item*2 at the same time. That is, a transaction with only *item*1 may be an outlier, but not a transaction with only *item*2. However, Fig. 1 shows that if the data point (*item*1, *item*2) is the center of a cluster and (*item*1) is an outlier, then (*item*2) should be an outlier, too. In other words, if transactions with only *item*1 are possible outliers, then transactions with only *item*2 should also be possible outliers according to distance. With higher dimensionalities more transactions will be incorrectly classified as outliers.

Many commercial applications rely on marketing transactional databases with both categorical and numerical attributes. Finding outlier transactions in such databases is important since outliers for instance may affect marketing management or sales strategies. Only a handful of studies have focused on the detection of outlier transactions from categorical datasets or transactional databases [15-16, 18, 21, 27]. He et al. [16] proposed an entropy-based method to detect outliers. The FindFPOF (Frequent Pattern Outlier Factor) algorithm [17] is another well-known item-based outlier detection technique. He, Xu and Deng [17] defined an outlier transaction as a transaction with few frequent patterns. FindFPOF first discovers frequent itemsets and then finds outliers by comparing each transaction with every frequent itemset. The drawback of this algorithm is that the efficiency deteriorates with the increase of frequent itemsets. Narita and Kitagawa [26] proposed another item-based approach where outliers are assumed to be transactions that violate most association rules. Narita and Kitagawa's work shared similarities with that of He, Xu and Deng [17], but they reduced the search space to expedite the search of outliers in large datasets.

Table 1. Transactional database sample 1. TID: Transaction IDentification number.

TID	Items
1	*item*1, *item*2
2	*item*1, *item*2
3	*item*1, *item*2

4	*item*1, *item*2
5	*item*1, *item*2
6	*item*1
7	*item*2
8	*item*2
9	*item*2
10	*item*2

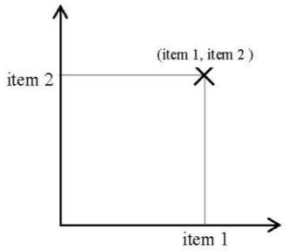

Fig. 1. The relationship between data point (*item*1, *item*2).

Secondly and most importantly, the attempts documented in the literature did not offer any suggestions on what caused the transactions to become abnormal.

In fact, the abnormal transactions themselves provide worthless information on decision making. For example, assume that an association rule *r*, {*Jam*, *Milk*}→{*Bread*}, is derived from a transactional database *D* with the high confidence value of 80%. According to the association rule *r*, the transaction <*Bacon*, *Corn*, *Jam*, *Milk*> may be abnormal due to the absence of *Bread*. There is no benefit of knowing the absence of a certain item due to its irrelevance. However, if one can find the reasons for occurrence of outliers that will help us make better decision.

There may be a variety of reasons behind the abnormal transactions. For example, a customer may want to buy bread, but found that he or she did not bring enough cash. In this example it is not easy to explore the underlying reasons for not buying bread. But, if the reason is because the emergence of some items leads to the disappearance of some other items, then finding the reason is transformed to the issue of identifying which items cause some other items' absence and this can be resolved by using the proposed method.

In the aforementioned example, could *Bacon* or *Corn* be the item that makes *Bread* absent? The idea to identify which items cause some other items' absence is practical. The reason is if users can apply association rules to find the relationships between items, they can also use association rules to find the relationships between items and absent items.

This paper proposes a framework for identifying the outlier transactions in marketing databases and finding which items may cause transactions to become outliers. The framework is divided into two parts.

The first part of this study is to utilize association rules to efficiently identify abnormal transactions in database. An abnormal transaction is defined as a transaction that is expected to contain some items that actually do not appear. Those items that should have been contained are marked as absent items. Absent items themselves hardly provide any value in decision making unless the reasons that cause the items' absence can be found.

The second part of this study uses association rules mining algorithm to extract the relationship between absent frequent items and infrequent items. Typically, the infrequent items that are always ignored in association rules study may be the key to cause transactions to be abnormal [13]. Our approach is to transform each transaction to absent frequent itemsets and infrequent items. These new transformed transaction can be mined by employing an association rules mining algorithm to find the relationship between infrequent items and absent items.

The remaining sections of this paper are organized as follows. Section 2 introduces related work on outlier detection. Section 3 describes the proposed method and section 4 identifies items that induce abnormal transactions. Section 5 provides experimental evidence. Section 6 concludes the paper.

2 Related work

2.1 Frequent itemsets and association rules

A frequent itemset is an itemset that contains a certain number of transactions. Association rules can be derived from frequent itemsets. The well-known association rule example derived from supermarket shopping data is {*diaper*}→{*beer*}, which means people buying diaper will also buy beer at the same time. The association rules help businesses to plan proper strategies to increase their sales. The following is a brief description of how association rules are found based on a transactional database.

Let $I = \{i_1, i_2, \ldots, i_m\}$ be the set of all items. A transactional database *D* is a set of transactions where each transaction *t* is a set of items such that $t \subseteq I$. The cardinality of the database *D* is denoted by |*D*|. For two itemsets $X, Y \subseteq I$ and $X \cap Y = \phi$, the rule $X \rightarrow Y$ means if *X* occurs then *Y* also occurs.

An itemset *X*'s support is denoted by *support(X)*:

$$support(X) = \frac{|X|}{|D|}. \quad (1)$$

An itemset is frequent if its support is larger than or equal to a pre-defined support threshold *min_sup*.

The confidence of $X \rightarrow Y$ is defined as *confidence(X→Y)*:

$$confidence(X \rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}. \quad (2)$$

An association rule is a rule with its confidence larger than or equal to a pre-defined threshold *min_conf*.

The Apriori-based algorithm is usually adopted for mining association rules. The original Apriori-based algorithm is inefficient because it repeatedly scans the same database to find frequent itemsets. Various non-Apriori methods have been proposed to expedite the discovery of association rules [14, 20, 33].

FP-growth [14], a well-known non-Apriori association rules mining algorithm, scans the database twice to build an FP-tree where all the frequent itemsets are stored. Each branch in the FP-tree is a frequent itemset. The association rules are then mined from the FP-tree. Since the FP-tree is a compact structure, its performance is better compared to the Apriori family of algorithms [11, 19].

2.2 Maximal frequent itemsets

The FP-tree structure is not only used to generate association rules but also a good data structure for applications that only need to utilize the information of frequent itemsets. However, if there are many long transaction patterns or the minimum support setting is low, the number of frequent itemsets and the FP-tree storage will be huge [11]. In this case, one can consider getting maximal frequent itemsets instead of frequent items.

A frequent itemset *X* is a maximal frequent itemset (MFI) if there is no other frequent itemset *Y* such that $X \subset Y$. Any subset of a maximum frequent itemset is a frequent itemset; that is, one still can get frequent itemsets from maximum frequent itemsets. Since the total number of maximum frequent itemsets is less than frequent itemsets, storage requirements are reduced. Several algorithms, such as MAFIA [5], GenMax [9] and FPmax [10], find maximal frequent itemsets.

FPmax is based on FP-growth and is proven to be a competitive algorithm [19]. FPmax builds an FP-tree like structure called an MFI-tree, to keep track of all maximum frequent itemsets. Subsequent research has proposed more effective algorithms for acquiring maximum frequent itemsets; however, since they are extensions of FP-growth and have huge storage requirements, the algorithm proposed herein employs FPmax to find maximum frequent itemsets. The following example illustrates how FPmax finds maximum frequent itemsets [10].

Table 2 lists a sample database that contains 10 transactions. The minimum support is set to be 20%. Fig. 2 shows the final complete FP-tree. If a FP-tree has only one path, it is a MFI-tree. Since the FP-tree in Fig. 2 is not a single path tree we first find a conditional pattern base and conditional FP-tree for each item in the header table. For example, the corresponding conditional FP-tree of item *f* is shown in Fig. 3. The items in the conditional pattern base are listed in descending order according to frequency. Note that if the conditional FP-tree of an item has more than one path the FP-tree needs to be separated into several single-path trees. The initial MFI-tree only contains the header table. The first item *f*'s conditional FP-tree is inserted into the MFI-tree and the itemset {*a, c, e, b, f*} is a maximum frequent itemset. The following step involves checking if the item *d*'s conditional FP-tree, {*a, c, d*}, is a subset of any maximum frequent itemset in the MFI-tree. If it is not a subset, it is inserted into the MFI-tree. Next, the item *b*'s conditional FP-tree, {a, c, e, b}, is a subset of {*a, c, e, b, f*} and will not be inserted. This subset-checking step is repeated until all the items in the header table are processed. Fig. 4 shows the complete MFI-tree.

Table 2. Transactional database sample 2.

TID	Items
1	*a, b, c, e, f, o*
2	*a, c, g*
3	*e, i*
4	*a, c, d, e, g*
5	*a, c, e, g, l*

6	*e, j*
7	*a, b, c, e, f, p*
8	*a, c, d*
9	*a, c, e, g, m*
10	*a, c, e, g, n*

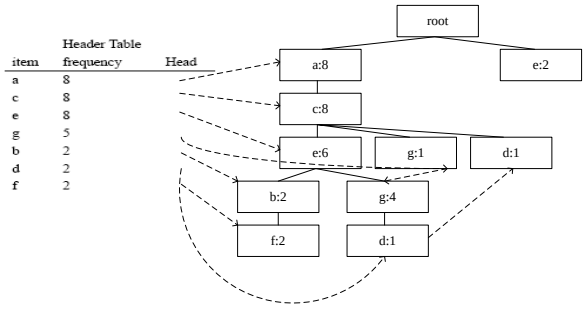

Fig. 2. The complete FP-tree derived from Table 2.

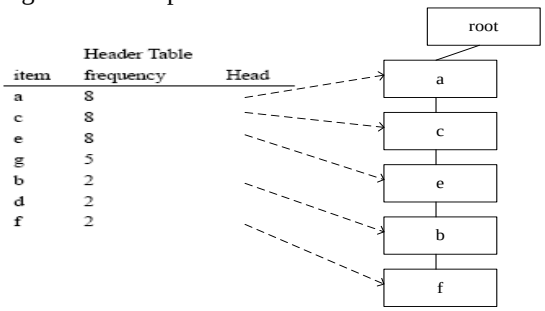

Fig. 3. The conditional FP-trees for item *f* in header table.

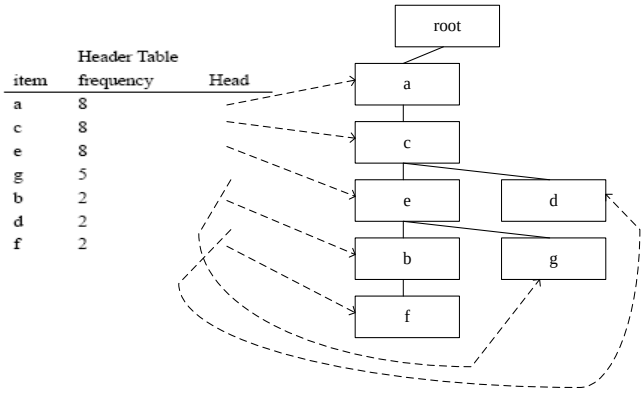

Fig. 4. The complete MFI-tree for the dataset in Table 2.

2.3 The definition of outlier transactions

There is no definite outlier definition. Different applications define outliers differently with different outlier detection approaches. Statistical methods fit the dataset to assumed distributions, and data are determined to be outliers according to how well they fit into the dataset [17]. However, the underlying distribution for a certain dataset may not match the assumed distribution and consequently affect the outlier detection accuracy. Another problem involves datasets with high dimensionality as it is difficult to estimate multidimensional distributions [22].

Distance-based methods define a data point *p* in a set *D* as an outlier if a certain percentage of other points in *D* are more than a pre-defined distance away from *p* [4]. Approaches using this definition [1, 34] have the drawback of high computation complexity when processing large datasets, making it difficult to find local outliers [22]. Clustering methods are also used to identify outliers, that is, points that are not inside any of the clusters [2, 3, 8, 24, 29]. Cluster-based methods first identify the clusters, thus the efficiency depends on how clusters are formed. Density-based methods [18] identify the outliers by comparing the density of the input dataset, and consider the outliers as points lying in low density regions.

All the mentioned methods are intended for datasets with numerical attributes and rely on data point distance measures to determine outliers.

Marketing transactional database, a multi-categorical attributes dataset, is usually employed to record data by commercial applications. In such multi-dimensional dataset, the concept of proximity may not be meaningful [23]. That is, an outlier transaction is not a data point; therefore, one cannot use the concept of distance measures to determine it.

The aforementioned methods are also not suitable to detect outlier transactions, even if some of the methods map the categorical attributes to numerical attributes before distances between data points are computed. The approach still faces the problem that the mapping results are not consistent across different mapping orderings. Only a few studies have focused on identifying outliers from transactional datasets [15-16, 27]. He et al. assumed that transactions containing less frequent itemsets are more likely to be outlier transactions [17]. They defined the Frequent Pattern Outlier Factor (FPOF) to evaluate whether a transaction is an outlier or not.

Narita and Kitagawa were interested in assessing if a transaction is likely to be an outlier when some items are supposed to appear, but actually do not appear [26]. Based on this concept, an outlier degree is defined to evaluate whether a single transaction is an outlier or not. According to their experiments Narita and Kitagawa claim that their approach can derive more accurate results compared to other approaches such as [17]. The following example illustrates the concept of outlier transactions [26].

In order to derive association rules from the transactional database in Table 3, the minimum support and minimum confidence are set to 50% and 80%, respectively. Table 4 gives partial association rules generated from Table 3. Since all the rules in Table 4 have high confidence, we see that TID 2 <*Bacon*, *Corn*, *Jam*, *Milk*> is abnormal. By checking RID 2, this transaction does not include the item *Bread* that is supposed to appear in the transaction. In fact, TID 2 is an outlier according to [26].

Table 3. Transactional database sample 3.

TID	Items
1	*Bread*, *Jam*, *Milk*
2	*Bacon*, *Corn*, *Jam*, *Milk*
3	*Bread*, *Jam*, *Milk*
4	*Bacon*, *Bread*, *Corn*, *Egg*, Milk
5	*Bacon*, *Bread*, *Corn*, *Egg*, *Jam*, *Milk*
6	*Bread*, *Corn*, *Jam*, *Milk*
7	*Bacon*, *Bread*, *Egg*, *Milk*
8	*Bacon*, *Bread*, *Egg*, *Jam*, *Milk*
9	*Bread*, *Jam*, *Milk*
10	*Bacon*, *Egg*, *Milk*

Table 4. Association rules derived from Table 3. RID: association Rules IDentification number.

RID	Rule
1	{*Jam*}→{*Bread*}
2	{*Jam*, *Milk*}→{*Bread*}
3	{*Jam*}→{*Bread*, *Milk*}
4	{*Bacon*}→{*Egg*}
5	{*Bacon*, *Milk*}→{*Egg*}
6	{*Bacon*}→{*Egg*, *Milk*}
7	{*Milk*}→{*Bread*}

2.4 The approach to detect abnormal transactions

This study aims to find the items which cause transactions to be abnormal. From the perspective of outlier management, we can approach from identifying the outlier transactions, analyzing the causality among items and then discovering the reasons behind the abnormality. Thus, our proposed abnormal transaction detection model will start from defining what abnormal transactions are. Then, we will propose our new finding on detecting abnormal transactions as well as on identifying items that induce abnormal transactions. The underlying section will introduce the definitions on deriving outlier degree.

Definition 1. Let t be a transaction, e be an item, and R be the set of high confidence association rules. t's associative closure t^+ is defined as follows:

$$t^0 = t$$
$$t^{i+1} = t^i \cup \left\{ e \mid e \in Y \text{ and } X \subseteq t^i \text{ and } X \to Y \in R \right\}$$
$$t^+ = t^\infty$$

The itemset t^{i+1} includes the item that should appear in t^i but actually does not. The itemset t^{i+1} will converge if t^i has no more items that should appear but actually do not appear. The associative closure t^+ is an ideal form for t and does not violate any association rule.

Definition 2. Let t be a transaction, R be the set of high confidence association rules, and t^+ be the associative closure of t. The outlier degree of t is defined as $od(t)$:

$$od(t) = \frac{\left| t^+ - t \right|}{\left| t^+ \right|}. \tag{3}$$

The outlier degree value is in the range 0 and 1. For example, the associative closure t^+ for TID 2 in Table 3 is <*Bacon, Corn, Bread, Egg, Jam, Milk*>. The outlier degree for TID 2 is therefore equal to $\frac{2}{6} = 0.33$. Note that if t^+ is equal to t, the outlier degree, $od(t)$, is equal to 0.

Definition 3. An outlier transaction is a transaction with an outlier degree $od(t)$ greater than or equal to *min_od*, a pre-defined outlier degree threshold.

If the *min_od* is set to 0.3, TID 2 in Table 3 is an outlier. Since the efficiency of the algorithm deteriorates with the increase of transactions' size, there is a need to improve the algorithm to reduce the time complexity. The basic idea is to reduce the size of both transactional database and the set of association rules.

Definition 4. Let M be the set of all maximal frequent itemset, m_i be a maximal frequent itemset and $m_i \in M$. A transaction t's maximal associative closure $t^+_{\max}$ is defined as follows:

$$t^0_{\max} = t$$
$$t^{i+1}_{\max} = t^i_{\max} \cup \left\{ e \mid e \in m_i \text{ and } m_i \cap t^i_{\max} \neq \phi \right\}$$
$$t^+_{\max} = t^\infty_{\max}$$

Definition 5. Let $od(t)$ be t's outlier degree, and $od(t)$'s upper bound is derived as follows:

$$od_{\max}(t) = \frac{\left| t^+_{\max} - t \right|}{\left| t^+_{\max} \right|}. \tag{4}$$

If the upper bound of a transaction's outlier degree is less than *min_od*, then the transaction is marked as an outlier. Instead of using associate rules set to identify outliers, one can first utilize maximal frequent itemsets with comparatively smaller data size to calculate each transaction's upper bound of outlier degree and then prune transactions with upper bounds less than *min_od*. This helps to reduce the transaction set, and the outlier degree is only computed for the remaining transactions. Consequently, the outlier degree calculation efficiency is significantly improved.

Definition 6. An association rule $X \to Y$ is a non-redundant rule if no other rules $Z \to W$ and $S \to V$ such that (*i*) $X \cup Y = Z \cup W$, $X \supset Z$, and (*ii*) $X = S$, $Y \subset V$, respectively.

According to Definition 6, RID 1, 2 and RID 4, 5 in Table 4 are redundant since they can be described by RID 3 and RID 6, respectively. The redundant rules can be removed from the original association rules set and the set of all non-redundant rules is denoted as the minimal rules set $R_{\min}$. The size of $R_{\min}$ is smaller than the original association rules set and a certain transaction's associative closure that derived from association rules set is the same as the associative closure derived from the minimal rules set. Definition 6 and RID rules given in Table 4 indicate that the efficiency of the outlier discovery algorithm is improved when the number of association rules is reduced.

3 Infrequent items and outlier degree

Outlier degree is a measurement to determine how many frequent items are absent in a specific transaction. It decides the completeness in detecting possible outlier transactions. By retrospect to Eq.(3), one can find that the infrequent items also affect the outlier degree calculation. We will show that there is no need to take infrequent items into consideration in calculating outlier degree.

Definition 7. Infrequent items are items that are not contained in any frequent itemsets.

The infrequent items in the transaction set in Table 5 are *Battery* and *Corn* using the same minimum support and minimum confidence as in the previous example. Compared with TID 2 in Table 3, this transaction has only one additional infrequent item, namely *Battery*. The associative closure for TID 2 is <*Bacon*, *Corn*, *Jam*, *Milk*, *Battery*, *Bread*, *Egg*>, and the outlier degree is equal to 2/7. If the minimum outlier degree is set to 0.3, this transaction is no longer an outlier. But according to [26], an outlier is a transaction with some items that are expected to appear, but do not. TID 2 in Table 5 is effectively the same as TID 2 in Table 3, and it should be an outlier. The problem arose from the definition of associative closure where the more infrequent items a transaction has, the more normal the transaction is. The outlier degree calculation on TID 2 and 10 in Table 5 are influenced by the number of infrequent items and there are no outliers if the minimum outlier degree is set to 0.3.

Based on this observation, one should remove infrequent items from the transactions before calculating the outlier degree. This discovery is based on the fact that the outlier degree is used to indicate how many frequent items are missing; therefore, the infrequent items should not be considered in calculating outlier degree.

Table 5. Transactional database sample 4.

TID	Items	Infrequent items	od(t)
1	*Bread, Jam, Milk*		0
2	*Battery, Bacon, Corn, Jam, Milk*	*Battery, Corn*	2/7
3	*Bread, Jam, Milk*		0
4	*Bacon, Bread, Corn, Egg, Milk*	*Corn*	0
5	*Bacon, Bread, Corn, Egg, Jam, Milk*	*Corn*	0
6	*Bread, Corn, Jam, Milk*	*Corn*	0
7	*Bacon, Bread, Egg, Milk*		0
8	*Bacon, Bread, Egg, Jam, Milk*		0
9	*Bread, Jam, Milk*		0
10	*Battery, Bacon, Egg, Milk*	*Battery*	1/5

Conventional method in calculating outlier degree cannot truly reflect the role of outliers in transactions. We therefore redefine a transaction's associative closure and its maximal associative closure to discover the outlier transactions.

Definition 8. Let t be a transaction, R be the set of high confidence association rules, and I_r be the set of all infrequent items. t^- is denoted as t's frequent transaction if all the infrequent items are removed from t. t's associative closure t^+ is defined as follows:

$$t^- = t - \{e \mid e \in I_r\}$$
$$t^0 = t^-$$
$$t^{i+1} = t^i \cup \{e \mid e \in Y \text{ and } X \subseteq t^i \text{ and } X \to Y \in R\}$$
$$t^+ = t^\infty$$

Definition 9. Let t be a transaction, R be the set of high confidence association rules, and t^+ be the associative closure of t. The new outlier degree of t should be defined as $od(t)$:

$$od(t) = \frac{\left|t^+ - t^-\right|}{\left|t^+\right|}. \tag{5}$$

Definition 10. Let M be the set of all maximal frequent itemset and $m_i \in M$. A transaction t's maximal associative closure $t^+_{\max}$ is defined as follows:

$$
\begin{aligned}
t^- &= t - \{e | e \in I_r\} \\
t^0_{\max} &= t^- \\
t^{i+1}_{\max} &= t^i_{\max} \cup \{e \mid e \in m_i \text{ and } m_i \cap t^i_{\max} \neq \phi\} \\
t^+_{\max} &= t^\infty_{\max}
\end{aligned}
$$

Definition 11. Let $od(t)$ be t's outlier degree, and $od(t)$'s upper bound is derived as follows:

$$od_{\max}(t) = \frac{\left|t^+_{\max} - t^-\right|}{\left|t^+_{\max}\right|}. \tag{6}$$

Fig. 5 shows the proposed outlier degree algorithm.

1.Get the association rules set R from a transactional database D by employing FP-growth algorithm.
2.Get the maximum frequent itemsets set M by employing FPmax algorithm.
3.Reduce the size of the transactional database.
  Get each transaction t's frequent transaction t^- and then calculate t's outlier degree upper bound od_{max}.
  Remove transactions whose od_{max} are less than *min_od*.
  The remaining transactions are the candidates of outlier transactions. The remaining transactions set is denoted as D_{min}.
4. Reduce the size of the association rules set.
  Remove redudant rules from R and get the minimum association rules set R_{min}.
5.Get the outlier transactions set OT.
  For each t in D_{min}
    Get each transaction t's frequent t^- and t's associative closure by checking R_{min}.
    Calculate t's outlier degree $od(t)$.
    If $od(t)$ >= *min_od* then $OT = OT \cup \{t\}$

Fig. 5. The proposed outlier degree algorithm.

4 Finding items that make transactions abnormal

The proposed outlier degree measurement method allows us to identify abnormal transactions. However, what is the benefit from discovering outliers? Can the discovery of outliers provide valuable information to further improve decision making? Usually, results from data mining help users realize unknown but important facts and users can utilize these facts to do some better decision making. For example, the famous association rule, {*diapers*} -> {*beers*}, mined from retail stores databases shows that those who purchase diapers tend to also buy beers when they go grocery shopping. Based on this observation, the retailers stock diapers next to the beer coolers to increase revenues. Intrusion detection, another outlier mining example, provides time series patterns to help users predict possible intrusion events. While an abnormal transaction may be detected, there is nothing we can do about it.

At first glance, it seems that the abnormal transactions themselves did not provide valuable information for knowledge mining and the proposed algorithm has no major improvement over the conventional methods. However, the major contribution of the presented work lies in finding the items that cause transactions to be abnormal. According to our knowledge there is no literature that studied on converting the outliers into useful knowledge. There could be thousands of reasons that cause transactions to be abnormal. Some reasons, like human errors, are not easy to predict and trying to explore

them is beyond the scope of this study. But, infrequent items may cause abnormal behavior in some applications [13] and we should go one step further to identify which items cause transaction to be abnormal.

To counterbalance this problem, a method is proposed for analyzing the relationships between infrequent items and abnormal transactions and identifying infrequent items that often cause certain frequent items' absence. Association rules mining finds items that are frequently occurring together. However, the mechanisms for finding association rules can also be applied to finding infrequent items that cause specific frequent items to be discarded. Before the association rules mining algorithm can be applied to find infrequent items that cause transactions to be labeled as abnormal, each transaction is transformed into two parts, namely absent frequent itemsets and infrequent items.

Definition 12. Let t be a transaction, $R_{\min}$ be the set of minimum association rules. An absent frequent itemset (*AF*) is defined as follows:

$$AF(t) = \left\{ e \mid e \in X \cup Y \text{ and } X \subseteq t \text{ and } Y \not\subset t \text{ and } X \to Y \in R_{\min} \right\}.$$

Table 6 lists the transformed transactional database from Table 5. There are no absent frequent itemsets and infrequent itemset in TID 1 in Table 5. TID 2 has three absent frequent itemsets, {*Milk*, *Bread**}, {*Milk*, *Jam*, *Bread**} and {*Bacon*, *Milk*, *Egg**}, an asterisk is used to denote that the item is expected, but actually does not occur. The infrequent items for the transaction are *Battery* and *Corn*. TID 4 has no absent frequent itemsets, but has one infrequent item *Corn*. TID 10 has one absent frequent itemset, {*Milk*, *Bread**}, and its infrequent itemset is *Battery*.

In the transformed transaction set, each absent frequent itemset is viewed as an item, and the relationships between absent frequent itemsets and infrequent items can be found. The complete algorithm including outlier detection and finding the relationship between infrequent items and outliers is shown in Fig. 6.

Table 7 shows an example used to verify that the proposed method can find the relationship between outlier and its infrequent items. The items in the synthetic transaction set are *a*, *b*, *c*, *d*, *e*, *f*, *g*, *h* and *i*. Table 8 shows partial frequent itemsets and part of association rules derived from Table 7 include $\{c\} \to \{d\}$, $\{d\} \to \{c\}$, $\{c, f\} \to \{d\}$, $\{d, f\} \to \{c\}$ if the minimum support is set to 50% and minimum confidence is 80%. According to Table 8, the infrequent items are *a*, *e*, *g*, *h*, and *i*. If the minimum outlier degree is set to 0.5, then transactions 5, 8, and 14 are outliers.

Table 6. The transformed transactional database. Each transaction is divided into unoberved frequent itemsets and infrequent itemsets.

TID	Items	Note
1	ϕ	ϕ denotes no absent frequent itemset and no infrequent item.
2	*Milk*/*Jam*/*Bread**, *Bacon*/*Milk*/*Egg**, *Milk*/*Bread**, *Battery*, *Corn*	The absent frequent itemset {*Milk*, *Jam*, *Bread**} is viewed as an item, and is denoted as *Milk*/*Jam*/*Bread**. *Battery* and *Corn* are infrequent items.
3	ϕ	
4	*Corn*	*Corn* is an infrequent item.
5	*Corn*	*Corn* is an infrequent item.
6	*Corn*	*Corn* is an infrequent item.
7	ϕ	
8	ϕ	
9	ϕ	
10	*Milk*/*Bread**, *Battery*	

1.Get the outlier transactions set *OT* from the transactional database *D*.
2.Transform *OT* to OT_{trans}
    Transform each *t* in *OT* to t_{trans} by dividing *t* into two parts, absent frequent itemsets and infrequent items.
3.Get the association rules set R_{trans} from OT_{trans}.

Fig. 6. Algorithm for finding abnormal transactions and identifying which items cause transaction to be labelled as abnormal.

The first step involves transforming each outlier into two parts, namely absent frequent itemsets and infrequent items. TID 5 has two absent frequent itemsets, {*d*, *c**} and {*d*, *f*, *c**}, and *e* and *h* are infrequent items. TID 8 has two absent frequent itemsets, namely {*c*, *d**} and {*c*, *f*, *d**}, and *e* and *i* are infrequent items. The final transformation result is shown in Table 9. Table 9 can be viewed as a new transactional database and each absent frequent itemset can be treated as an item. By applying association rules mining with minimum support and minimum confidence set to 50% and 80%, respectively, we find that item *h* is the one that induces the abnormal transaction since rule $\{h\} \rightarrow \{d, f, c^*\}$ and $\{h\} \rightarrow \{d, c^*\}$ can be derived from Table 9. It means that item *c* should appear, but because of the infrequent item *h*, item *c* is not observed in the transaction. That is, item *h* causes the transaction to be marked as abnormal.

Table 7. A partial synthetic transactional database.

TID	Items
1	*c, d, f, g*
2	*a, b, c, d, e, g*
3	*a, c, d, f*
4	*c, d, h, i*
5	*d, e, f, h*
6	*a, c, d, f, e, g*
7	*b, c, d, e, f*
8	*b, c, f, e, i*
9	*c, d, e, f, g, i*
10	*b, c, d, f*
11	*a, b, c, d*
12	*b, g*
13	*c, d, f, h*
14	*b, d, f, h*
15	*b, c, d, f*
16	*c, d, f, g*

Table 8. Part of frequent itemsets and infrequent items derived from Table 7.

1-item frequent itemsets	2-item frequent itemsets	3-item frequent itemsets	infrequent items
{*b*}	{*f, c*}	{*f, c, d*}	*a, e, g, h, i*
{*c*}	{*f, d*}		
{*d*}	{*c, d*}		
{*f*}			

Table 9. The transformed transactional database according to Table 8.

TID	Items
5	*d*/*c**, *d*/*f*/*c**, *e*, *h*
8	*c*/*d**, *c*/*f*/*d**, *e*, *i*
14	*d*/*c**, *d*/*f*/*c**, *h*

5 Experimental results and discussion

Three experiments were conducted to evaluate the effectiveness of the algorithm. The proposed algorithm was implemented in Dev C++ and experiments were run on a workstation with an Intel 2.5GHz processor and 2G of memory. The FP-growth is adopted to mine frequent itemsets and association rules. The maximum frequent itemsets are derived by using FPmax.

The first experiment uses a synthetic data set as input generated using IBM Quest synthetic data generator. The parameter settings for the data generation are: (*i*) the total number of transactions $|D|$=532, (*ii*) average size per transaction $|t|$=8, and (*iii*) total number of items $|N|$=25. To get association rules from the generated 532-transactional database, minimum support and minimum confidence were set to 18% and 78%, respectively. Table 10 lists the discovered association rules and infrequent items. Before getting outliers, the redundant association rules check was performed and no redundant rules were found in this experiment. If the minimum outlier degree is set to 10%, 96 outliers are detected by employing conventional algorithm [26], compared to 106 outliers by employing the approach proposed herein. Table 11 shows the number of outliers detected with different minimum outlier degree settings. The proposed algorithm can prune more transactions and find more outliers than the conventional methods. It is important to find any possible outliers that may induce the transactions to be abnormal.

Next, the 106 outliers were taken as the testing set to find which items cause the outliers to be marked as abnormal. First, each outlier was transformed into absent frequent itemsets and infrequent items, and then the association rules mining was applied to the transformed set. Table 12 shows partial transformed result. In order to discover the items that cause the transactions to be marked as abnormal, several settings were explored and it was found that with minimum support and minimum confidence being set to 5% and 50%, respectively, the rule {*j*} -> {*f, i, m**} is found. This means that the infrequent item *j* causes transactions to be marked as abnormal.

Table 10. Association rules derived from the 532 transactions generated by data generator.

Association rules	Infrequent items
{*d, i*}→{*m*}	*h, j, l, n, o, q, t, u, x*
{*d, v*}→{*m*}	
{*f, b*}→{*m*}	
{*f, i*}→{*m*}	
{*b, c*}→{*m*}	
{*b, v*}→{*m*}	
{*b, i, v*}→{*m*}	
{*b, i, m*}→{*v*}	

Table 11. Outliers discovered from the transactional database.

	No. of transactions pruned		No. of outliers discovered	
min. outlier degree	previous [26]	our	previous [26]	our
10%	136	141	96	106
20%	175	188	33	45

Table 12. Partial transformed data set for the 106 outliers.

TID	Items
9	*b*/*i*/*m*/*v**, *j*, *u*
26	*d*/*v*/*m**, *j*
39	*b*/*i*/*m*/*v**
89	*d*/*i*/*m**, *d*/*v*/*m**, *b*/*v*/*m**, *b*/*i*/*v*/*m**, *t*
	…

The second experiment uses Wisconsin breast cancer data set from UCI Machine Learning Repository [37]. In order to check the efficiency, accuracy and precision rates are defined as follows:

$$accuracy = \frac{\text{no. of detected outliers that are positive}}{\text{no. of all outliers}}. \tag{7}$$

$$precision = \frac{\text{no. of detected outliers that are positive}}{\text{no. of detected outliers}}. \tag{8}$$

The original Wisconsin breast cancer data set contains 699 records with 458 labeled as benign and 241 labeled as malignant. Each record has 9 attributes and one class attribute. The attribute information is shown in Table 13. Among the

699 records, 14 benign records and 2 malignant records containing unknown data are discarded. To form an unbalanced data set, the experiment follows the strategy outlined [12], namely removing another 200 malignant records. The final test data contains 444 benign records and 39 malignant records. We assume the 39 malignant records are true abnormal records.

We also assume that some attributes may cause certain records to be abnormal. In order to derive relationships between attributes, each record is transformed into a multi-categorical attributes' transaction and then the association rules algorithm can be applied to this transaction set to get association rules. For example, if the first attribute value is 5, it will be labeled as *a*5. If the second attribute value is 1, it will be labeled as *b*1 (see Table 14). The transaction with class attribute *o*2 is a benign record, whereas *o*4 is a malignant record.

The third step involves mining association rules from the test data. Since the goal is to detect malignant records that are thought as outliers, only rules with consequent part *o*2 are kept. Table 15 shows the association rules discovered from the transformed data set with minimum support and confidence set to 75% and 85%, respectively. Note that according to Definition 6 all *k*-item rules with *k* greater than 2 are redundant and are not listed in Table 15. The top-*k* highest outlier degrees are chosen as outliers, that is, an outlier is not decided by comparing its outlier degree with the pre-defined minimum outlier degree. Table 16 lists top-10, top-20, top-40, and top-60 true outlier number detected with corresponding accuracy and precision rate. According to Table 16, the proposed algorithm yields better accuracy and precision rates than previous approach.

The last step of the second experiment involves finding infrequent items that cause outliers to be marked as abnormal. The experiment tries to find items in the top-40 result. Again, each outlier in the top-40 result is transformed into absent frequent itemsets and infrequent items. The minimum support and minimum confidence are set to 50% and 80%, respectively. Only rule {*f*a} -> {*b*1, *o*2*} is found. That is, the item *f*a (the attribute Bare Nuclei is 10) may be the reason that caused a patient's tumor to be malignant, although the attribute Uniformity of Cell Size is 1.

The third experiment uses Parkinson's telemonitoring data set from UCI Machine Learning Repository [37]. There are a total of 5,875 records in the data set, and each record has 19 attributes capturing 16 voice measures, gender, motor-UPDRS score, and total-UPDRS score [31]. Each attribute is quantitative and needs to be discretized, or divided into several intervals, before the association rules mining algorithm is applied. Each attribute, except gender, is divided into three non-intersected intervals, high, medium and low. For example, age ranges from 36 to 85, and a subject older than 74 belongs to the high interval, below 50 belongs to the low and others belong to the medium range.

After discretizing the attributes, one can proceed to discover the association rules from the Parkinson's data set. In this experiment, a record with motor-UPDRS_medium is assumed to be a normal record while a record with motor-UPDRS_high is treated as a possible outlier record.

Similar to the second experiment, only rules with consequent part of motor-UPDRS_medium are kept. Table 17 shows partial association rules discovered from the transformed data set with minimum support and confidence of 8% and 75%, respectively. No k-item rule with k less than 4 has consequent part of motor-UPDRS_medium, and according to Definition 6, all k-item rules with k greater than 4 that have consequent part of motor-UPDRS_medium are redundant. Table 18 lists discovered infrequent items.

The second step involves finding outliers by comparing the Parkinson's data set with the discovered rules. Several possible outlier records are found and if the minimum outlier degree is set to 0.05, the 4 records in Table 19 will be true outliers. These 4 records should have attribute value motor-UPDRS_medium according to the association rules discovered, but they have motor-UPDRS_high. To find infrequent items that cause the records to become abnormal, each outlier in Table 19 is transformed into two parts, absent frequent itemsets and infrequent items. Again, by applying the association rules mining algorithm with a minimum support of 50% and a minimum confidence of 50%, the infrequent item RPDE_low (a nonlinear dynamical complexity measure below 0.347) is identified as the source that causes the records to become abnormal. That is, when all the measurements are in medium or low intervals, the RPDE measure is the key to tell a healthy subject apart from a Parkinson's patient.

These experiments show that the proposed outlier detection method is more practical than previous approaches since it not only identifies the outlier transactions, but also discovers the associations between outlier and its infrequent items. This is useful since the mining results help users determine what causes transactions to become abnormal without having consulted expert knowledge in advance.

Table 13. The attribute information for Wisconsin breast cancer data set.

attribute ID	attribute information	domain
1	Clump Thickness	1-10
2	Uniformity of Cell Size	1-10
3	Uniformity of Cell Shape	1-10
4	Marginal Adhesion	1-10
5	Single Epithelial Cell Size	1-10
6	Bare Nuclei	1-10

7	Bland Chromatin	1-10
8	Normal Nucleoli	1-10
9	Mitoses	1-10
10	Class	2 for benign, 4 for malignant

Table 14. Partial transformed data set in transaction format.

Original instances (9 categorical attributes in each record)	Transformed result
<5, 1, 1, 1, 2, 1, 3, 1, 1, 2>	<*a*5, *b*1, *c*1, *d*1, *e*2, *f*1, *g*3, *h*1, *i*1, *o*2>
<5, 4, 4, 5, 7, 10, 3, 2, 1, 2>	<*a*5, *b*4, *c*4, *d*5, *e*7, *f*a, *g*3, *h*2, *i*1, *o*2>
<3, 1, 1, 1, 2, 2, 3, 1, 1, 2>	<*a*3, *b*1, *c*1, *d*1, *e*2, *f*2, *g*3, *h*1, *i*1, *o*2>
<9, 1, 2, 6, 4, 10, 7, 7, 2, 4>	<*a*9, *b*1, *c*2, *d*6, *e*4, *f*a, *g*7, *h*7, *i*2, *o*4>
…	…

Table 15. Association rules mined from Wisconsin breast cancer data set.

2-item rules (support, confidence)
{*b*1} -> {*o*2} (76.4%, 100%)
{*f*1} -> {*o*2} (80.1%, 98.7%)
{*h*1} -> {*o*2} (81.0%, 98.5%)
{*i*1} -> {*o*2} (89.2%, 95.1%)

Table 16. Outlier dection under different k values.

Top-k	previous (accuracy, precision) [26]	our (accuracy, precision)
top-10	6 (15%, 60%)	8 (21%, 80%)
top-20	15 (38%, 75%)	17 (44%, 85%)
top-40	34 (87%, 85%)	35 (90%, 88%)
top-60	39 (100%, 65%)	39 (100%, 65%)

Table 17. Partital association rules discovered from Parkinson's data set (Min Support=8%, Min Confidence=75%).

Antecedent part for discovered rule	Consequent part for discovered rule
Male, shimmer_low, dB_low	motor-UPDRS_medium
Male, shimmer_low, APQ3_low	motor-UPDRS_medium
Male, age_medium, dB_low	motor-UPDRS_medium
Male, dB_low, RPDE_medium	motor-UPDRS_medium
Male, APQ3_low, DFA_medium	motor-UPDRS_medium
…	…

Table 18. Infrequent items discovered from Parkinson's data set (Min Support=8%, Min Confidence=75%).

age_low, jitter_high, Abs_high, RAP_high, PPQ5_high, DDP_high, shimmer_high, dB_high, APQ3_high,APQ5_high, APQ11_high, DDA_high, NHR_high, HNR_low, HNR_high, RPDE_low, RPDE_high, DFA_high, PPE_high

Table 19. The 4 outliers discovered by comparing the Parkinson's data set with discovered association rules.

No.	Outlier records	Outlier degrees
1	age_low, Male, motor-UPDRS_high, jitter_low, Abs_medium, RAP_low, PPQ5_low, DDP_low, shimmer_low, dB_low, APQ3_low, APQ5_low, lN,DDA_low, NHR_low, HNR_medium, RPDE_medium, DFA_medium, PPE_medium	1/19
2	age_medium, Male, motor-UPDRS_high, jitter_low, Abs_low, RAP_low, PPQ5_low, DDP_low, shimmer_low, dB_low, APQ3_low, APQ5_low, APQ11_low, DDA_low, NHR_low, HNR_high, RPDE_medium, DFA_low, PPE_low	1/19
3	age_medium, Male, motor-UPDRS_high, jitter_low, Abs_low, RAP_low, PPQ5_low, DDP_low, shimmer_low, dB_low, APQ3_low, APQ5_low, APQ11_low, DDA_low, NHR_low, HNR_medium, RPDE_low, DFA_low, PPE_medium	1/19
4	age_medium, Male, motor-UPDRS_high, jitter_low, Abs_low, RAP_low, PPQ5_low, DDP_low, shimmer_low, dB_low, APQ3_low, APQ5_low, APQ11_low, DDA_low, NHR_low, HNR_medium, RPDE_low, DFA_low, PPE_medium	1/19

6 Conclusions

From the perspective of outlier management, conventional methods did not tackle the question on how to further utilize the detected outliers. The proposed framework can find the infrequent items that induce the transactions to be abnormal. To prevent the infrequent items from deviating from the true outlier degrees the proposed method modified the definition of transaction's association closure by removing the infrequent items before the calculation of outlier degrees.

After identifying the outliers, the proposed approach further discovers which infrequent items make transactions abnormal. Abnormal transactions are divided into absent frequent itemsets and infrequent items. By applying association rule mining method, the relationship between absent frequent itemsets and infrequent items are found. Items that cause the transactions to become outliers are therefore found and the mining results are easier to understand. The proposed framework provides a total solution not only on finding but also on managing outliers. The experimental results verify that the proposed algorithm is more efficient both in terms of accuracy and precision rates.

Future improvements are possible. The calculation of outlier degree relies on associative closure. However, the confidence values of association rules should be considered. That is, if a transaction violates a higher confidence rule, it should have higher outlier degree. Next, the precedent parts of association rules affect the outlier detection. Since the proposed algorithm employs non-redundant rules to check transactions, the final result may include many known outliers, and even the infrequent items that cause abnormal outliers are revealed. In this case, setting a minimum item number for precedent part may solve the problem.

We first apply the framework to health care data to verify the algorithm's feasibility. In the future, it is necessary to acquire more real world data from different sources to derive abnormal transactions and find reasons behind the abnormality. The mining results will also be shared with hospital officials to inquire their opinions. It is important to mention that the proposed framework can also be applied to any kind of transaction data set to find which infrequent items induce transactions to be abnormal. There are a variety of reasons that can lead to abnormality, and this study's contribution is to provide a way to identify the sources of confusion. The infrequent items are always ignored in data mining but now they may provide valuable information to allow people to make better decision.

Acknowledgments

This work was supported in part by Ministry of Science and Technology, Taiwan under Grants NSC102-2221-E-027-083- and NSC102-2218-E-002-009-MY2, and in part by joint project between National Taipei University of Technology and Mackay Memorial Hospital under Grant NTUT-MMH-102-03 and Grant NTUT-MMH-103-01.

References

[1] Angiulli F, Pizzuti C (2002) Fast outlier detection in high dimensional spaces. In: Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery in Databases. Helsinki, Finland: 15-26.

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

[3] Bahrampour S, Moshiri B, Salahshoor K (2011) Weighted and constrained possibilistic C-means clustering for online fault detection and isolation. Applied Intelligence 35(2): 269-284.

[4] Bhaduri K, Matthews BL, Giannella CR (2011) Algorithms for speeding up distance-based outlier detection. In: Proceedings of ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, San Diego, CA, USA: 859-867.

[5] Burdick D, Calimlim M, Flannick J, Gehrke J, Yiu T (2005) MAFIA: A maximal frequent itemset algorithm. IEEE Trans on Knowledge and Data Engineering 17:1490-1504.

[6] Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: A survey. ACM Computing Surveys 41:1-58.
[7] Chazard E, Ficheur G, Bernonville S, Luyckx M, Beuscart R (2011) Data mining to generate adverse drug events detection rules. IEEE Trans on Information Technology in Biomedicine 15:823-830.
[8] Elahi M, Li K, Nisar W, Lv X, Wang H (2008) Efficient clustering-based outlier detection algorithm for dynamic data stream. In: Proceedings of the 5th Int. Conf. on Fuzzy Systems and Knowledge Discovery, Jinan, Shandong, China 5:298-304.
[9] Gouda K, Zaki MJ (2001) Efficiently mining maximal frequent itemsets. In: Proceedings of IEEE Int. Conf. on Data Mining, San Jose, California, USA: 163-170.
[10] Grahne G, Zhu J (2003) High performance mining of maximal frequent itemsets. In: Proceedings of the 6th SIAM Workshop on High Performance Data Mining, San Francisco, CA, USA: 135-143.
[11] Grahne G, Zhu JF (2005) Fast algorithms for frequent item set mining using FP-Trees. IEEE Trans on Knowledge and Data Engineering 17:1347-1362.
[12] Guo T, Li GY (2008) Neural data mining for credit card fraud detection. In: Proceedings of the 7th Int. Conf. on Machine Learning and Cybernetics, Kunming, China 7: 3630-3634..
[13] Haglin DJ, Manning AM (2007) On minimal infrequent itemset mining. In: Proceedings of the Int. Conf. on Data Mining, Las Vegas, Nevada, USA: 141-147.
[14] Han J, Pei J, Yin Y (2000) Mining frequent patterns without candidate generation. In: Proceedings of ACM SIGMOD Int. Conf. on Management of Data, Dallas, Texas, USA: 1-12.
[15] He Z, Deng S, Xu X (2005) An optimization model for outlier detection in categorical data. In: Proceedings of IEEE Int. Conf. on Intelligent Computing, Hefei, China: 400-409.
[16] He Z, Deng S, Xu X (2006) A fast greedy algorithm for outlier mining. In: Proceedings of the 10th Pacific-Asia Conf. on Knowledge Discovery and Data Mining, Singapore: 567-576.
[17] He Z, Xu X, Deng S (2005) Fp-outlier: Frequent pattern based outlier detection. Computer Science and Information System 2:103-118.
[18] Hido S, Tsuboi Y, Kashima H, Sugiyama M, Kanamori T (2011) Statistical outlier detection using direct density ratio estimation. Knowledge and Information Systems 26:309-336.
[19] Hu T, Sung SY, Xiong H, Fu Q (2008) Discovery of maximum length frequent itemsets. Information Sciences 178: 69-87.
[20] Huang Y-P, Kao LJ, Sandnes FE (2008) Efficient mining of salinity and temperature association rules from ARGO data. Expert Systems with Applications 35:59-68.
[21] Koufakou A, Georgiopoulos M, Anagnostopoulos GC, Reynolds KM (2007) A scalable and efficient outlier detection strategy for categorical data. In: Proceedings of IEEE Int. Conf. on Tools with Artificial Intelligence, Patras, Greece: 210-217.
[22] Koufakou A, Georgiopoulos M (2010) A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes. Data Mining and Knowledge Discovery 20: 259-289.
[23] Kriegel HP, Kröger P, Zimek A (2009) Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Transactions on Knowledge Discovery from Data 3:1-58.
[24] Lei D, Zhu QH, Chen J, Lin H, Yang P (2012) Automatic PAM clustering algorithm for outlier detection. Journal of Software 7:1045-1051.
[25] Márquez-Vera C, Morales CR, Soto SV (2013) Predicting school failure and dropout by using data mining techniques. IEEE Journal of Latin-American Learning Technologies 8:7-14.
[26] Narita K, Kitagawa H (2008) Outlier detection for transaction databases using association rules. In: Proceedings of the 9th Int. Conf. on Web-Age Information Management, Zhangjiajie, Hunan, China: 373-380.
[27] Otey ME, Ghoting A, Parthasarathy A (2006) Fast distributed outlier detection in mixed-attribute data sets. Data Mining and Knnowledge Discovery 12:203-228.
[28] Papadimitriou S, Kitagawa H, Gibbons PB, Faloutsos C (2003) Loci: Fast outlier detection using the local correlation integral. In: Proceedings of the 19th Int. Conf. on Data Engineering, Bangalore, India: 315-326.
[29] Shi K, Li L (2013) High performance genetic algorithm based text clustering using parts of speech and outlier elimination. Applied Intelligence 38(4): 511-519.
[30] Troiano L, Scibelli G (2014) Mining frequent itemsets in data streams within a time horizon. Data & Knowledge Engineering 89:21-37.
[31] Tsanas A, Little MA, McSharry PE, Ramig LO (2010) Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. IEEE Transactions on Biomedical Engineering 57:884-893.
[32] Tseng VS, Shie B-E, Wu C-W, Yu PS (2013) Efficient algorithms for mining high utility itemsets from transactional databases. IEEE Trans on Knowledge and Data Engineering 25:1772-1786.
[33] Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan G, Ng A, Liu B, Yu P, Zhou Z-H, Steinbach M, Hand D, Steinberg D (2008) Top 10 algorithms in data mining. Knowledge and Information Systems 14:1-37.
[34] Wan Y, Bian F (2008) Cell-based outlier detection algorithm: A fast outlier detection algorithm for large datasets. In: Proceedings of the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Osaka, Japan 5012:1042-1048.
[35] Yanqing J, Hao Y, Peter D, Ayman M, John T, Richard ME, Massanari R-M (2011) A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. IEEE Trans on Information Technology in Biomedicine 15:428-437.
[36] Zhu C, Kitagawa H, Faloutsos C (2005) Example-based robust outlier detection in high dimensional datasets. In: Proceedings of the 5th IEEE Int. Conf. on Data Mining, Houston, Texas, USA: 829-832.
[37] UCI machine learning repository. http://www.ics.uci.edu/˜mlearn/MLRepository.html.