Mining Correlated Patterns with Multiple Minimum All-Confidence Thresholds

R. Uday kiran and Masaru Kitsuregawa

Institute of Industrial Science, The University of Tokyo, Komaba-ku, Tokyo, Japan.
uday_rage@tkl.iis.u-tokyo.ac.jp and kitsure@tkl.iis.u-tokyo.ac.jp.

Abstract. Correlated patterns are an important class of regularities that exist in a database. The *all-confidence* measure has been widely used to discover the patterns in real-world applications. This paper theoretically analyzes the all-confidence measure, and shows that, although the measure satisfies the null-invariant property, mining correlated patterns involving both frequent and rare items with a single minimum all-confidence (*minAllConf*) threshold value causes the "rare item problem" if the items' frequencies in a database vary widely. The problem involves either finding very short length correlated patterns involving rare items at a high *minAllConf* threshold, or generating a huge number of patterns at a low *minAllConf* threshold. The cause for the problem is that the single *minAllConf* threshold was not sufficient to capture the items' frequencies in a database effectively. The paper also introduces an alternative model of correlated patterns using the concept of multiple *minAllConf* thresholds. The proposed model facilitates the user to specify a different *minAllConf* threshold for each pattern to reflect the varied frequencies of items within it. Experiment results show that the proposed model is very effective.

Keywords: Knowledge discovery, frequent patterns, correlated patterns and *rare item problem*.

1 Introduction

1.1 Background and Related Work

Mining frequent patterns (or itemsets) [1] from transactional databases has been actively and widely studied in data mining. In the basic model, a pattern is said to be frequent if it satisfies the user-defined minimum support (*minSup*) threshold. The *minSup* threshold controls the minimum number of transactions that a pattern must cover in a transactional database. Since only a single *minSup* threshold is used for the entire database, the model implicitly assumes that all items within a database have uniform frequencies. However, this is often not the case in many real-world applications. In many applications, some items appear very frequently in the data, while others rarely appear. If the items' frequencies vary a great deal, mining frequent patterns with a single *minSup* threshold leads to the dilemma known as the *rare item problem* [2]. It involves either completely missing of frequent patterns involving rare items at a high *minSup* threshold or generating too many patterns at a low *minSup* threshold.

To confront the rare item problem in applications, alternative interestingness measures have been discussed in the literature [3,4]. Each measure has a selection bias that justifies the significance of a knowledge pattern. As a result, there exists no universally acceptable best measure to judge the interestingness of a pattern for any given dataset or application. Researchers are making efforts to suggest a right measure depending upon the user and/or application requirements [5–7].

Recently, *all-confidence* is emerging as a measure that can disclose true correlation relationships among the items within a pattern [8,9]. The two key reasons for its popular adoption are *anti-monotonic* and *null-invariant properties*. The former property facilitates in the reduction of search space as all non-empty subsets of a correlated pattern must also be correlated. The latter property facilitates to disclose genuine correlation relationships without being influenced by the object co-absence in a database. The basic model of correlated patterns is as follows [10].

Let $I = \{i_1, i_2, \cdots, i_n\}$ be a set of items, and DB be a database that consists of a set of transactions. Each transaction T contains a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called TID. Let $X \subseteq I$ be a set of items, referred as an itemset or a *pattern*. A pattern that contains k items is a k-pattern. A transaction T is said to contain X if and only if $X \subseteq T$. The support of a pattern X in DB, denoted as $S(X)$, is the number of transactions in DB containing X. The pattern X is said to be **frequent** if it occurs no less frequent than the user-defined minimum support (*minSup*) threshold, i.e., $S(X) \geq minSup$. The *all-confidence* of a pattern X, denoted as *all-conf(X)*, can be expressed as the ratio of its support to the maximum support of an item within it. That is, $all\text{-}conf(X) = \frac{S(X)}{max\{S(i_j) | \forall\ i_j \in X\}}$.

Definition 1. ***(The correlated pattern** X.) The pattern X is said to be* **all-confident** *or* **associated** *or* **correlated** *if*

$$\begin{aligned} S(X) &\geq minSup \\ &and \\ \frac{S(X)}{max(S(i_j) | \forall i_j \in X)} &\geq minAllConf. \end{aligned} \tag{1}$$

Where, *minAllConf* is the user-defined minimum all-confidence threshold value. The *minSup* threshold controls the minimum number of transactions a pattern must cover in a database. The $minAllConf$ threshold controls the minimum number of transactions a pattern must cover with respect to its items' frequencies in a database.

Table 1: Transactional database.

TID	ITEMS	TID	ITEMS	TID	ITEMS	TID	ITEMS	TID	ITEMS
1	a, b	5	c, d	9	c, d, g	13	a, b, e	17	c, d
2	a, b, e	6	a, c	10	a, b	14	b, e, f, g	18	a, c
3	c, d	7	a, b	11	a, b	15	c, d	19	a, b, h
4	e, f	8	e, f	12	a, c, f	16	a, b	20	c, d, f

Example 1. Consider the transactional database of 20 transactions shown in Table 1. The set of items $I = \{a,\ b,\ c,\ d,\ e,\ f,\ g,\ h\}$. The set of items a and b, i.e., $\{a,\ b\}$ is a pattern. It is a 2-pattern. For simplicity, we write this pattern as "*ab*". It occurs in 8 transactions (*tids* of $1,2,7,10,11,13,16$ and 19). Therefore, the support of "*ab*," i.e., $S(ab) = 8$. If the user-specified $minSup = 3$, then "*ab*" is a frequent pattern because $S(ab) \geq minSup$. The *all-confidence* of "*ab*", i.e., $all\text{-}conf(ab) = \frac{8}{max(11,9)} = 0.72$. If the user-specified $minAllConf = 0.63$, then "*ab*" is a correlated pattern because $S(ab) \geq minSup$ and $all\text{-}conf(ab) \geq minAllConf$.

1.2 Motivation

A pattern-growth algorithm known as CoMine has been proposed in [10] to discover the complete set of correlated patterns. We introduced the concept of items' support intervals and proposed an improved CoMine algorithm known as CoMine++ [11]. While testing the performance of our algorithm on various datasets, we have observed that although the null-invariant property of all-confidence facilitates the effective discovery of correlated patterns involving both frequent and rare items, the usage of a single $minAllConf$ threshold for the entire database confines the effective discovery of correlated patterns involving both frequent and rare items to the databases in which the items' frequencies do not vary widely. If the items' frequencies in the database vary widely, then mining correlated patterns with a single $minAllConf$ threshold can lead to the following problems:

- At a high $minAllConf$ threshold, many discovered correlated patterns involving rare items can have very short length. Most of them were singleton patterns as they always have $all\text{-}conf = 1$.
- In order to discover long correlated patterns involving rare items, we have to set low $minAllConf$ threshold. However, it causes combinatorial explosion producing too many correlated patterns.

We call this dilemma as the "rare item problem". In this paper, we analyze the all-confidence measure, introduce concepts to describe the problem and propose a generalized correlated pattern model to address the problem.

1.3 Contributions of this Paper

i. In this paper, we theoretically analyze the all-confidence measure and show that mining correlated patterns involving both frequent and rare items with a single $minAllConf$ leads to the *rare item problem*.

ii. We describe the reason for the problem by using the concepts known as *items' support intervals* and *cutoff-item-supports*.

iii. We also propose a technique to confront the problem. The technique is based on the notion of multiple constraints, where each pattern can satisfy a different $minAllConf$ value depending upon the frequencies of items within itself.

iv. By conducting experiments on various datasets, we show that the proposed technique can discover interesting correlated patterns involving both frequent and rare items without generating a huge number of meaningless correlated patterns.

1.4 Paper Organization

The rest of this paper is organized as follows. Section 2 introduces the rare item problem in the correlated pattern model. Section 3 describes the proposed model. Section 4 presents the experimental evaluations of basic and proposed models. Finally, Section 5 concludes the paper with future research directions.

2 The Rare Item Problem in Correlated Pattern Mining

In this section, we first introduce the concepts "items' support intervals" and "cutoff-item-support." These two concepts facilitate us to understand the problem, which is discussed subsequently.

2.1 Items' Support Intervals

The concept of *items' support intervals* was introduced in [11]. It says that every item can generate correlated patterns of higher-order by combining with only those items that have support within a specific interval. For the user-defined $minAllConf$ and $minSup$ thresholds, the support interval of an item $i_j \in I$ is $\left[max\left(\begin{matrix} S(i_j) \times minAllConf, \\ minSup \end{matrix}\right), max\left(\begin{matrix} \frac{S(i_j)}{minAllConf}, \\ minSup \end{matrix}\right)\right]$. The correctness is given in [11]. The support interval width of items plays a key role in correlated pattern mining. It defines the range of items' frequencies with which an item can combine to generate correlated patterns of higher-order.

Example 2. The support of f in Table 1 is 5. If $minAllConf = 0.63$ and $minSup = 3$, then the item 'f' can generate correlated patterns of higher order by combining with only those items that have support in the range of [3, 8] $(= [max(5 \times 0.63, 3), max(\frac{5}{0.63}, 3)])$.

If the support interval width of an item is large, then it can combine with wide range of items' supports to generate correlated patterns of higher order. If the support interval width of an item is relatively small, then it can combine with only a small range of items' supports to generate correlated patterns of higher order.

2.2 Cutoff-item-support

The concept of items' support intervals alone is not sufficient to describe the rare item problem. Therefore, we introduce another concept, called *cutoff-item-support*, to describe the problem. It is as follows.

From the definition of correlated pattern (see Equation 1), it turns out that if X is a correlated pattern, then

$$S(X) \geq max(max(S(i_j)|\forall i_j \in X) \times minAllConf, minSup) \tag{2}$$

Based on Equation 2, we introduce the following definition.

Definition 2. *(Cutoff-item-Support of an item.) The* cutoff-item-support *(CIS) for an item $i_j \in I$, denoted as $CIS(i_j)$, is the minimum support a pattern $X \ni i_j$ must have to be a correlated pattern. It is equal to the maximum of minSup and product between its support and minAllConf. That is, $CIS(i_j) = max(minSup,\ S(i_j) \times minAllConf)$.*

Example 3. The support of '*a*' in Table 1 is 11. If the user-specified $minSup = 3$ and $minAllConf = 0.63$, then the *CIS* of '*a*', denoted as $CIS(a) = 7$ ($\simeq max(3,\ 0.63 \times 11)$). It means any pattern containing '*a*' must have support no less than 7 to be a correlated pattern. Similarly, the *CIS* values of '*b*', '*c*', '*d*', '*e*', '*f*', '*g*' and '*h*' are 6, 6, 4, 3, 3, 3 and 3, respectively.

The *CIS* of an item plays a key role in correlated pattern mining. If the items have their *CIS* values very close to their supports, then it is difficult to expect long patterns containing those items (due to the apriori property [1]). If the items have their *CIS* values very far away from their supports, then it is possible to discover long patterns involving those items. However, this can cause combinatorial explosion, producing too many patterns as the items can combine with one another in all possible ways. Thus, while specifying the *minAllConf* and *minSup* thresholds, we have to take care that items' *CIS* values are neither too close nor too far away from their respective supports.

Definition 3. *(Redefinition of a correlated pattern using items' CIS values.) A pattern X is said to be correlated if its support is no less than the maximum CIS value of all its items. That is, if X is a correlated pattern satisfying the user-defined minSup and minAllConf thresholds, then $S(X) \geq max(CIS(i_j)|\forall i_j \in X)$.*

Since the *CIS* of an item $i_j \in X$, i.e., $CIS(i_j) = max(S(i_j) \times minAllConf,\ minSup)$, the correctness of the above definition is straight forward to prove from Equation 2.

2.3 The Problem

When we use a single *minAllConf* threshold value to mine correlated patterns, then:

- The support interval width of items do not remain uniform. Instead it decreases from frequent items to (less frequent or) rare items (see Property 1). It is illustrated in Example 4.

  Example 4. In Table 1, the item '*a*' appears more frequently than the item '*d*'. If the user-defined $minSup = 3$ and $minAllConf = 0.63$, then the support intervals for '*a*' and '*d*' are $[7,17]$ and $[6,14]$, respectively. It can be observed that the support interval width of '*a*' is 10, while for '*f*' is only 8. Thus, the support interval decreases from frequent to rare items.

  Thus, the usage of a single *minAllConf* threshold facilitates only the frequent items to combine with wide range of items' supports.
- The difference (or gap) between the items' support and corresponding *CIS* values do not remain uniform. Instead, it decreases from frequent to rare items (see Property 2).

Example 5. Continuing with Example 3, it can be observed that the difference between the *support* and corresponding *CIS* values of '*a*' and '*d*' are 4 and 2, respectively. Thus, the gap between the items' support and corresponding *CIS* values decreases from frequent to rare items.

Property 1. Let i_j and i_k, $1 \leq j \leq n$, $1 \leq k \leq n$ and $j \neq k$, be the items such that $S(i_j) \geq S(i_k)$. For the user-defined *minAllConf* and *minSup* thresholds, it turns out that the support interval width of i_j will be no less than the support interval width of i_k. That is,

$$max\begin{pmatrix} S(i_j) \times minAllConf, \\ minSup \end{pmatrix} - max\begin{pmatrix} \dfrac{S(i_j)}{minAllConf}, \\ minSup \end{pmatrix} \geq$$
$$max\begin{pmatrix} S(i_k) \times minAllConf, \\ minSup \end{pmatrix} - max\begin{pmatrix} \dfrac{S(i_k)}{minAllConf}, \\ minSup \end{pmatrix} \quad (3)$$

Property 2. Let i_p and i_q be the items having supports such that $S(i_p) \geq S(i_q)$. For the user-specified *minAllConf* and *minSup* thresholds, it turns out that $S(i_p) - CIS(i_p) \geq S(i_q) - CIS(i_q)$.

The non-uniform width of items' support intervals and the non-uniform gap between the items' support and corresponding *CIS* values causes the following problems while mining correlated patterns with a single *minAllConf* threshold:

i. At a high *minAllConf* value, rare items will have *CIS* values very close (or almost equivalent) to their respective supports. In addition, the support interval width for rare items will be very small allowing only rare items to combine with one another. Since it is difficult for the rare items to combine with one another and have support which is almost equivalent to their actual support, many discovered rare correlated patterns may have very short length. We have observed that many correlated patterns involving rare items discovered at a high *minAllConf* threshold value were singleton patterns as they always have *all-conf* $= 1$.

ii. To discover long correlated patterns involving rare items, we must specify a low *minAllConf* threshold value. However, a low *minAllConf* causes frequent items to have their *CIS* values far away from their respective supports causing combinatorial explosion and producing too many correlated patterns.

We call this dilemma as the *rare item problem*. In the next section, we discuss the technique to address the problem.

3 Proposed Model

To address the *rare item problem*, it is necessary for the correlated pattern model to simultaneously set high *minAllConf* threshold for a pattern containing frequent items and low *minAllConf* threshold for a pattern containing rare items. To do so, we exploit the notion of "multiple constraints" and introduce the model of mining correlated patterns using multiple minimum all-confidence thresholds. The idea is to specify *minAllConf* threshold for each item depending upon its frequency (or occurrence behavior in the database) and specify the *minAllConf* for a pattern accordingly. We inherit this approach from the multiple *minSups*-based frequent pattern mining, where each pattern can satisfy a different *minSup* depending upon the items' frequencies within itself [12].

In our proposed model, the definition of correlated pattern remains the same. However, the definition of *minAllConf* is changed to address the problem. In the proposed model, each item in the database is specified with a *minAllConf* constraint, called *minimum item all-confidence* (*MIAC*). Next, *minAllConf* of the pattern $X = \{i_1, i_2, \cdots, i_k\}$, $1 \leq k \leq n$, is represented as maximum *MIAC* value among all its items. That is,

$$minAllConf(X) = max\begin{pmatrix} MIAC(i_1), & MIAC(i_2), \\ \cdots, & MIAC(i_k) \end{pmatrix} \tag{4}$$

where, $MIAC(i_j)$ is the user-specified *MIAC* threshold for the item $i_j \in X$. Thus, the definition of correlated pattern is as follows.

Definition 4. *(Correlated pattern.) The pattern* $X = \{i_1, i_2, \cdots, i_k\}$, $1 \leq k \leq n$, *is said to be correlated if*

$$\begin{array}{rcl} S(X) & \geq & minSup \\ & and & \\ all\text{-}conf(X) & \geq & max\begin{pmatrix} MIAC(i_1), & MIAC(i_2), \\ \cdots, & MIAC(i_k) \end{pmatrix} \end{array} \tag{5}$$

Please note that the correctness of *all-confidence* measure will be lost if *minAllConf* **of a pattern is represented with the *minimal*** *MIAC* **of all the items within itself.** In particular, if the items' *MIAC* values are converted into corresponding *CIS* values, then the above definition still preserves the definition of correlated pattern given in Definition 3.

Definition 5. *(Problem Definition.) Given the transactional database (DB), the minimum support (minSup) threshold and the items' minimum item all-confidence (MIAC) threshold values, discover the complete set of correlated patterns that satisfy the minSup and maximum MIAC value of all items within itself.*

The proposed model facilitates the user to specify the all-confidence thresholds such that the support interval width of items can be uniform and the gap between the items' supports and corresponding *CIS* values can also be uniform. In addition, it also allows the user to dynamically set high *minAllConf* value for the patterns containing frequent items and low *minAllConf* value for the patterns containing only rare items. As a result, this model can efficiently address the *rare item problem*. Moreover, the proposed model is the generalization of basic model to discover the patterns with all-confidence and support measures. If all items are specified with a same *MIAC* value, then the proposed model is same as the basic model of correlated pattern.

The correlated patterns discovered with the proposed model satisfy the *downward closure property*. That is, all non-empty subsets of a correlated pattern must also be correlated patterns. The correctness is based on Property 3 and shown in Lemma 1.

Property 3. If $Y \supset X$, then $minAllConf(Y) \geq minAllConf(X)$ as $max(MIAC(i_j)|\forall i_j \in Y) \geq max(MIAC(i_j)|\forall i_j \in X)$.

Lemma 1. *The correlated patterns discovered with the proposed model satisfy the downward closure property.*

Proof. Let X and Y be the two patterns in a transactional database such that $X \subset Y$. From the *apriori property* [1], it turns out that $S(X) \geq S(Y)$ and $all\text{-}conf(X) \geq all\text{-}conf(Y)$. Further, $minAllConf(Y) \geq minAllConf(X)$ (see, Property 3). Therefore, if $all\text{-}conf(Y) \geq minAllConf(Y)$, then $all\text{-}conf(X) \geq minAllConf(X)$.

Pei et al. [13] have proposed a generalized pattern-growth algorithm known as FIC^A (Mining frequent itemsets with convertible anti-monotone constraint) to discover patterns if they satisfy downward closure property. Since the patterns discovered with the proposed model satisfy the downward closure property, FIC^A algorithm can be extended to mine correlated patterns with the proposed model.

4 Experimental Results

In this section, we evaluate the basic and proposed models of correlated patterns. We show that proposed model allows us to find correlated patterns involving rare items without generating a huge number of meaningless correlated patterns with frequent items.

Since the FIC^A algorithm discussed in [13] can discover the complete set of patterns for any measure that satisfies the downward closure property, we extend it to discover correlated patterns with either single *minAllConf* or multiple *minAllConf* thresholds. We do not report the performance of FIC^A algorithm. They are available at [13].

4.1 Experimental setup

The FIC^A algorithm was written in GNU C++ and run with Ubuntu 10.04 operating system on a 2.66 GHz machine with 1GB memory. The runtime specifies the total execution time, i.e., CPU and I/Os. We pursued experiments on synthetic (T10I4D100K) and real-world (Retail [14] and BMS-WebView-2 [15]) datasets. The T10I4D100K dataset contains 100,000 transactions with 870 items. The Retail dataset contains 88,162 transactions with 16,470 items. The BMS-WebView-2 dataset contains 77,512 transactions with 33,401 items. All of these datasets are sparse datasets with widely varying items' frequencies.

4.2 A method to specify items' *MIAC* values

For our experiments, we need a method to assign *MIAC* values to items in a dataset. Since the non-uniform difference (or gap) between the items' support and *CIS* values is cause of *rare item problem*, we employ a method that specifies items' *MIAC* values such that there exists a uniform gap between the items' *support* and *CIS* values. It is as follows.

Let δ be the representative (e.g. mean, median, mode) support of all items in the database. Choosing a $minAllConf$ value, calculate the gap between support and corresponding CIS values (see Equation 6). We call the gap as "support difference" and is denoted as SD.

$$\begin{aligned} SD &= \delta - \delta \times minAllConf \\ &= \delta \times (1 - minAllConf) \end{aligned} \tag{6}$$

Next, we specify $MIAC$ values for other items such that the gap (or SD) remains constant. The methodology to specify $MIAC$ values to the items $i_j \in I$ is shown in Equation 7.

$$MIAC(i_j) = max\left(\left(1 - \frac{SD}{S(i_j)}\right),\ LMAC\right) \tag{7}$$

Where, $LMAC$ is the user-defined lowest $MIAC$ value an item can have. It is particularly necessary as $\left(1 - \frac{SD}{S(i_j)}\right)$ can give negative $MIAC$ value for an item i_j if $SD > S(i_j)$.

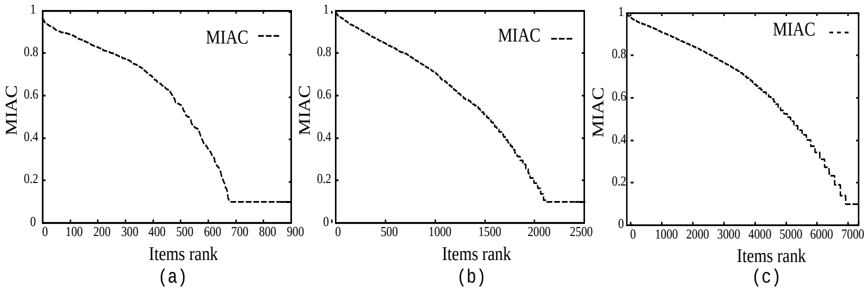


Fig. 1: The items' $MIAC$ values in different datasets. (a) T10I4D100k dataset (b) BMS-WebView-2 dataset and (c) Retail dataset.

Fig. 1(a), (b) and (c) respectively shows the $MIAC$ values specified by the proposed model in $T10I4D100K$, *BMS-WebView-2* and *Retail* datasets with $minAllConf = 0.9$, $LMAC = 0.1$ and δ representing the mean support of frequent items. The X-axis represents the rank of items provided in the descending order of their support values. The Y-axis represents the $MIAC$ values of items. It can be observed that the $MIAC$ values decreases as we move from frequent to less frequent (or rare) items. The low $MIAC$ values for rare items facilitate them to combine with other items and generate correlated patterns. The high $MIAC$ values of frequent items will prevent them from combining with one another in all possible ways and generating huge number of uninteresting patterns.

4.3 Performance results

Fig. 2(a), (b) and (c) respectively shows the number of correlated patterns generated in the basic model and proposed model of correlated patterns in T10I4D100K, BMS-WebView-2 and Retail datasets. The X-axis represents the $minAllConf$ values that are

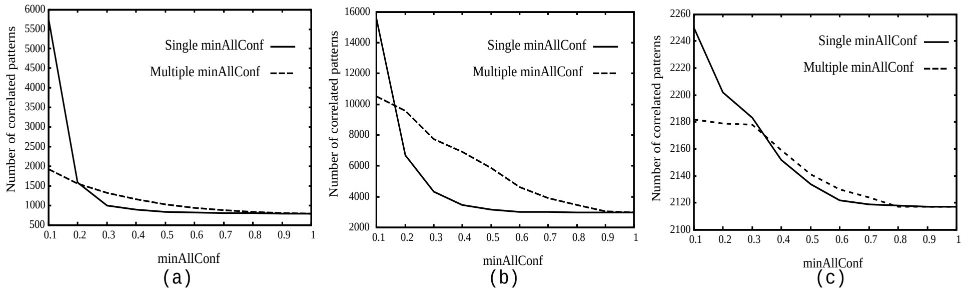


Fig. 2: Generation of Correlated patterns in different datasets. (a) T10I4D100k dataset (b) BMS-WebView-2 dataset and (c) Retail datasets.

varied from 0.1 to 1. The Y-axis represents the number of correlated patterns discovered at different *minAllConf* values. Too many correlated patterns were discovered in both models when $minAllConf = 0$. For convenience, we are not showing the correlated patterns discovered at $minAllConf = 0$. The following observations can be drawn from these two graphs:

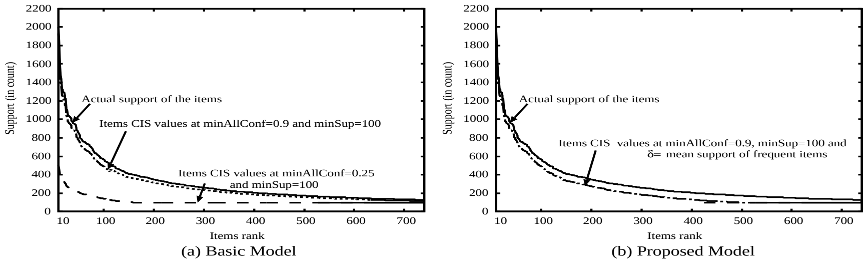
i. At high *minAllConf* values, the proposed model has generated more number of correlated patterns than the basic model. The reason is as follows. In the basic model, *CIS* values of rare items were very close (almost equal) to their respective supports. Since it is difficult for the rare items to combine with other (rare) items and have support as their own, very few correlated patterns pertaining to rare items have been discovered. In the proposed model, some of the rare items had their *CIS* values relatively away from their supports. This facilitated the rare items to combine with other rare items and generate correlated patterns.

ii. At low *minAllConf* values, the proposed model has generated less number of correlated patterns than the basic model. The reason is as follows. The basic model has specified low *CIS* values for the frequent items at low *minAllConf*. The low *CIS* values of the frequent items facilitated them to combine with one another in all possible ways and generate too many correlated patterns. In the proposed model, the *CIS* values for the frequent items were not very far away from their respective supports (as compared with basic model). As a result, the proposed model was able to prevent frequent items to combine with one another in all possible ways and generate too many correlated patterns.

4.4 A case study using BMS-WebView-2 Dataset

In this study, we show how the proposed model allows the user to find correlated patterns containing rare items without generating a huge number of meaningless correlated patterns with frequent items. Due to page limitations, we confine our discussion only to BMS-WebView-2 dataset. Similar observations were also drawn from the other datasets.

Fig. 3 (a) shows the *CIS* values specified for the items by the basic model. The following observations can be drawn from this figure. (*i*) The gap between the support and CIS values of items decreases as we move from frequent to rare items, irrespective

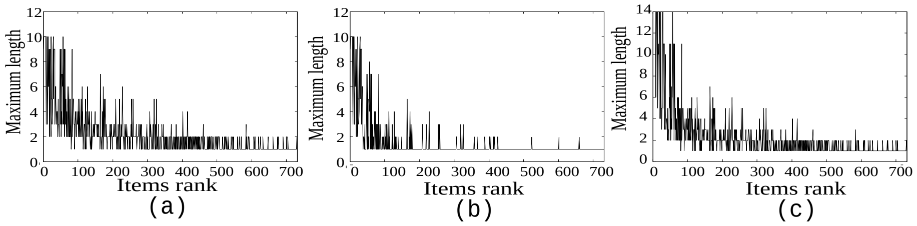


(a) Basic Model (b) Proposed Model

Fig. 3: The *CIS* values specified for the items by both the models. The *minSup* is expressed in support count.

of the *minAllConf* threshold value. (*ii*) At a high *minAllConf* value, rare items have the *CIS* values very close (or almost equivalent) to their respective supports. (*iii*) Choosing a low *minAllConf* value facilitates rare items to have *CIS* values relatively away from their supports. However, it causes frequent items to have their *CIS* values far away from their supports, which can cause combinatorial explosion.

Fig. 3 (b) shows the *CIS* values specified for items by the proposed model. It can be observed that by using item specific *minAllConf* values, appropriate gap between the *support* and *CIS* values of items can be maintained irrespective of their frequencies.



(a) (b) (c)

Fig. 4: Maximum length of the items in BMS-WebView-2 dataset. (a) Proposed model with high all-confidence threshold value. (b) Basic model with high all-confidence threshold value. (c) Basic model with low all-confidence threshold value.

Fig. 4(a) shows the maximum length of a correlated patten formed for each frequent item in the proposed model with $minSup = 100$ (in support count), $minAllConf = 0.9$ and δ representing the mean of frequent items. Fig. 4(b) provides the same information for the basic model with $minSup = 100$ and $minAllConf = 0.9$. In the basic model, majority of the rare items (i.e., items having rank greater than 400) had the maximal length of 1. In other words, correlated patterns involving rare items were only singleton patterns. In our proposed model, maximal length of correlated patterns involving rare items ranged from 1 to 4. This shows that the proposed model has facilitated the generation of correlated patterns containing both frequent and rare items at high *minAllConf* value.

At $minAllConf = 0.4$, the basic model has also facilitated the generation of correlated patterns containing rare items. However, the usage of low *minAllConf* threshold has caused combinatorial explosion producing too many correlated patterns. In Fig. 4(c), the increased maximal length of frequent items (i.e., items having rank from 1 to 100) from 10 (at $minAllConf = 0.9$) to 14 (at $minAllConf = 0.4$) signifies the combinatorial explosion in which frequent items have combined with one another in many possible ways.

5 Conclusion and Future Work

This paper has shown that although the all-confidence measure satisfies the null-invariant property, mining correlated patterns with a single *minAllConf* threshold in the databases of widely varying items' frequencies can cause the rare item problem. In addition, a generalized model of mining the patterns with the multiple *minAllConf* values has been introduced to confront the problem. The effectiveness of the new model is shown experimentally and practically.

As a part of future work, we would like to investigate the methodologies to specify items *MIAC* values. It is also interesting to investigate efficient mining of closed and maximal correlated patterns with the proposed model.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: SIGMOD, pp. 207–216 (1993)
2. Weiss, G. M.: Mining with rarity: a unifying framework. ACM SIGKDD Explorations Newsletter, Vol. 6, Issue 1, pp. 7–19 (2004)
3. Omiecinski, E. R.: Alternative interest measures for mining associations in databases. IEEE Trans. on Knowl. and Data Eng., Vol. 15, pp. 57–69 (2003)
4. Brin, S., Motwani, R., Silverstein, C.: Beyond market baskets: generalizing association rules to correlations. SIGMOD Rec. Vol. 26, pp. 265–276 (1997)
5. Tan, P. N., Kumar, V., Srivasta, J.: Selecting the right interestingness measure for association patterns. In: KDD, pp. 32–41 (2002)
6. Wu, T., Chen, Y., Han, J.: Re-examination of interestingness measures in pattern mining: a unified framework. Data Mining Knolwedge Discovery, Vol. 21, pp. 371–397 (2010)
7. Surana, A., Kiran, R. U., Reddy, P. K.: Selecting a Right Interestingness Measure for Rare Association Rules. In: COMAD, pp. 115–124 (2010)
8. Kim, W. Y., Lee, Y. K., Han, J.: Ccmine: efficient mining of confidence-closed correlated patterns. In: PAKDD, pp. 569–579 (2004)
9. Kim, S., Barsky, M., Han, J.: Efficient mining of top correlated patterns based on null invariant measures. In: ECML PKDD, pp. 172–192 (2011)
10. Lee, Y. K., Kim, W. Y., Cao, D., Han, J.: CoMine: efficient mining of correlated patterns. In: ICDM, pp. 581–584 (2003)
11. Kiran, R. U., Kitsuregawa, M.: Efficient Discovery of Correlated Patterns in Transactional Databases Using Items' Support Intervals. In: DEXA, pp. DEXA (1) 234–248 (2012)
12. Liu, B., Hsu, W., Ma, Y.: Mining association rules with multiple minimum supports. In: KDD, pp. 337–341 (1999)
13. Pei, J., Han, J., Lakshmanan, L. V.: Pushing convertible constraints in frequent itemset mining. Data Mining and Knowledge Discovery, Vol. 8, pp. 227–251 (2004)
14. Brijs, T., Goethals, B., Swinnen, G., Vanhoof, K., Wets, G.: A data mining framework for optimal product selection in retail supermarket data: the generalized PROFSET model. In: KDD, pp.300–304 (2000)
15. Zheng, Z., Kohavi, R., Mason, L.: Real world performance of association rule algorithms. In: KDD, pp. 401–406 (2001)