Mining conditional discriminative pattern

Xiaoqing Liu*, Jun Wu*, Haipeng Gong*, Shengchun Deng†, Zengyou He*

Abstract

here is abstract ,which I have to write

Key words: Discriminative pattern, contrast sets, emerging patterns, frequent pattern, data mining.

*Xiaoqing Liu, Jun Wu, Haipeng Gong and Zengyou He are with the School of Software, Dalian University of Technology, Dalian, China. *Email: eileenwelldone@gmail.com, wujun.myway@gmail.com, haipengxf@gmail.com, zyhe@dlut.edu.cn.*

†Shengchun Deng is with the School of Computer Science and Engineering, Harbin Institute of Technology, Harbin, China. *Email: dsc@hit.edu.cn.*

1 Introduction

Discriminative pattern mining aims at finding patterns that occur with disproportionate frequency in data sets with different class labels. These patterns are defined as discriminative patterns here although they may have some other names such as emerging patterns [4] and contrast sets [23]. For human beings, where the sex can be interpreted as labels, the amount of estrogen is absolutely different between male and female persons. So hormone can be characterized as discriminative pattern for certain. In fact, those biological entities such as differentially expressed genes or proteins in patients and healthy persons, are also examples of discriminative patterns. They exactly capture the differences between labeled data sets, which contributes to building high quality classfiers and characterizing different classes.

The discovery of discriminative patterns is of considerable value in many domains such as patient risk group detection in medicine, discovery of over-expressed genes in microarray data analysis, identification of distinguishing features in customer relationship management [19].

The algorithms for discovering discriminative patterns have been widely studied in different areas. In terms of terminology and task definitions, discriminative patterns can also be refined into contrast sets [23], emerging patterns [4] and subgroups [12],[26]. According to [22],[11],[23], contrast set mining aims at discovering patterns that capture prominent frequency differences across different user-defined groups of subjects. Emerging pattern mining detects patterns that capture frequency growth change from one class to another [4],[15]. While subgroup discovery tries to find population subgroups that are large enough and statistically "most interesting" [12],[26]. These methods simply appear with different names in distinct definitions, and they share in common as owners of discriminative power in essence. Accordingly, the algorithms for mining discriminative patterns can be categorized with respect to the discriminative measures and pattern searching strategies.

In perspective of measures for discriminative power of a pattern, discriminative pattern discovery has employed different measures such as *DiffSup* [6],[23],[14],[20], *GrowthRate* [4] and *WRAcc* [13],[3]. Notably, *DiffSup* of a pattern is defined as the difference of the supports between two classes. Here the support is defined as the percentage of instances in the corresponding class that contain this pattern. *DiffSup* is proposed in [23] and further extended in [6],[14],[20]. *GrowthRate* of a pattern from a pair of classes A_1 and A_2 is defined as the ratio of its support in class A_1 to that in class A_2. It presents the frequency increase of a pattern from one class to another. *WRAcc* is defined as follows:

$$WRAcc(pattern, class) = p(pattern) \cdot (p(class \mid pattern) - p(class)),$$

where p(pattern) stands for the occurrence probability of one pattern in all instances, and p(class) stands for the prior probability that samples belong to the current class. Similarly, p(class | pattern) stands for the conditional probability that the pattern is in the current class. Moreover, some other common measures can also achieve the same goal such as information gain [1], odds ratio [24], support ratio [4], etc.

Existing algorithms adopt different search strategies for fast discriminative pattern mining. Some algorithms make use of Apriori framework for this task [16],[17],[2],[27],[10]. However, most discriminative measures are not anti-monotonic [23],[1],[4]. To address this issue, some methods mine discriminative patterns without using the Apriori framework. One strategy is to conduct the mining process as a two-step procedure. It first generates all frequent patterns as candidates from one class, and later uses discriminative measures to check their discriminative power. Another strategy is to directly use a discriminative measure for pruning candidate patterns under the risk of missing some discriminative patterns..

Despite of the significant progress achieved during the past decades, there are still some issues that remain unsolved. One challenging problem is how to efficiently discover highly discriminative patterns with very low support values. To address this issue, we have to make a trade-off between the completeness and running efficiency. On one hand, some methods use a very low support

threshold to catch low-support discriminative patterns at the cost of much higher running time. On the other hand, researchers adopt some anti-monotonic measures such that we can find a larger fraction of low-support discriminative patterns within an acceptable amount of time [23]. Nevertheless, the second strategy will miss some target patterns as well.

Even the algorithms discussed above can generate a comprehensive set of discriminative patterns, these results may involve some biased patterns whose discriminative power mainly comes from its subsets. That is, the discriminative power of some patterns just benefits from their subsets instead of the whole body of themselves. So eliminating those uninteresting biased patterns is an essential work. To remove these redundant discriminative patterns, one method is to apply the permutation test in statistics [9] to post-process the mining results. However, as pointed out in [18], the standard permutation test procedure may not fully address this issue. As a result, they propose the "sequential permutation test" method that is capable of reducing the effect of subpatterns when calculating the statistical significance of discriminative patterns. Such sequential permutation test procedure is very effective in reducing the redundancy of reported discriminative patterns and is applied to different applications [7]. However, the permutation test process is very time-consuming in practice. Another alternative strategy is to choose a rigorous measure that can remove the influence of subsets of patterns. Existing methods such as Motif-X [21] and MMFPh [25] are examples of this idea.

In this paper, we propose the problem of conditional discriminative pattern mining by formalizing and generalizing previous research efforts in bioinformatics applications [21],[25]. We introduce the notion of 'conditional' to broaden the applicability of these existing measures with respect to subsets interplays. Owing to the problem formulation, the patterns reported would be strictly significant discriminative positives other than due to their subsets. In such a scenario, we make it possible to guarantee the non-redundancy of discriminative pattern discovery.

In the following, we will first illustrate the basic idea of conditional discriminative pattern and then show that this concept can remove patterns whose discriminative power is mainly due to their subsets. One pattern P is said to be a k-pattern if it has k items. If another pattern Q contains only a subset of these k items in P, then Q is a sub-pattern of P. In fact, every instance contains the pattern must also contain its corresponding sub-patterns, but that is not true in turn. So the set of instances that contain one pattern must be a subset of instance collections that contain its sub-pattern. Notably there are exactly k sub-patterns of size k-1 for one k-pattern. For each sub-pattern of size k-1, we can generate a subset of instances in which every instance contains this sub-pattern. On this new data set, we can re-calculate the statistical significance of the k-pattern. In this setting, this k-pattern is claimed as a conditional discriminative pattern if it is a discriminative pattern on all k sub-pattern induced data sets. Conditional discriminative pattern is defined with the set of instances induced from its sub-patterns as the background. As a result, the effects of its sub-patterns in the evaluation of discriminative power are reduced in an elegant manner.

To further illustrate how the concept of conditional discriminative pattern can filter out uninteresting patterns, we use the data in Figure 1 as an example. Here rows stand for items and columns for instances from two classes. We select 15 items and 20 instances to make up four different typological patterns, where P1 = { i_1, i_2, i_3 }, P2 = { i_5, i_6, i_7 }, P3 = { i_9, i_{10}} and P4 = { i_{12}, i_{13}, i_{14} }, respectively. For illustration purpose, we take *DiffSup* (the absolute support difference between two classes) as the discriminative measure.

In this example, P2 is obviously not a discriminative pattern since it has similar frequency between the two labeled classes. Thus, it cannot be a conditional discriminative pattern as well. In contrast, P1, P3 and P4 are discriminative patterns with lager significance values. In such a scenario, further discussion will be placed on these three patterns. P1 is a pattern that contains 3 items, which has three sub-patterns of length 2: {i_1, i_2}, {i_2, i_3}, and {i_1, ,i_3}. The set of instances containing {i_1, i_2} in class A is {5,6,7,8,9,10}, and that in class B is {11,12}. On this new sub-pattern induced data set, the *DiffSup* value for P1 is 1.0. Similarly, the *DiffSup* values for P1 on data sets induced from another two sub-patterns are 1.0 as well. As a result, P1 is definitely

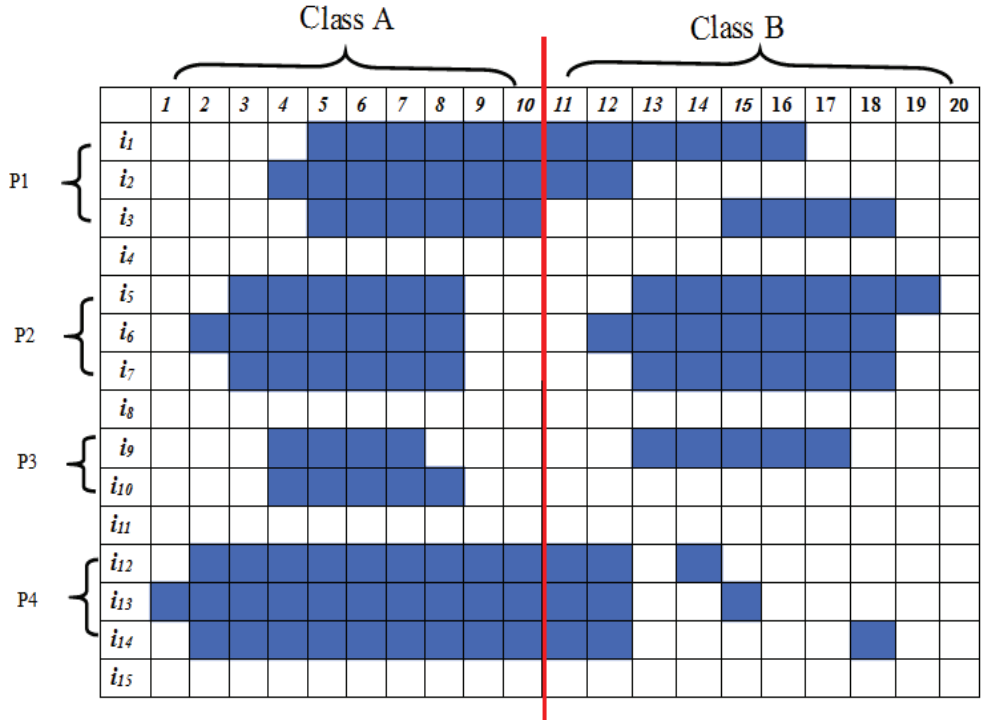

Figure 1: An sample data set containing twenty instances divided into two classes, fifteen items making up four different categorical patterns. Lines mean items, and columns indicate instances. Cells in shade represent current items exist in the corresponding instances and vice versa. For example, cell (i_1, 5) is in shade, so the fifth instance in Class A contains item i_1.

a conditional discriminative pattern with no matter which significance threshold is used since the *DiffSup* value is always less than or equal to 1.0. In the same way, the *DiffSup* values of P3 on its sub-pattern induced data sets are 1 and 0.17, respectively. As a result, P3 is not a conditional discriminative pattern if the significance threshold is larger than 0.3. In fact, patterns like P3 have been investigated in [5] and this type of patterns are defined as driver-passenger patterns according to the contrary interaction between subsets. Obviously, the discriminative power of P3 mainly comes from the contribution of i_9, and patterns of this type are not interesting in discriminative pattern discovery. Therefore, we can filter out these biased discriminative patterns according to our rigorous definition. Patterns like P4 are not conditional discriminative patterns as well.

To address the problem of conditional discriminative pattern mining, we present here two new algorithms, TFAD and TFBD, respectively. TFAD utilizes the support constraint together with discriminative power measures in the same mining process. While TFBD adopts a two-step strategy that we employ the support constraint to generate and prune candidates in a level-wise manner first, and then refine them using the discriminative measures. Experiments on real data sets show that these two algorithms are efficient and effective in conditional discriminative pattern mining.

The remainder of this paper is organized as follows: Section 2 presents the details of TFAD and TFBD algorithms. Section 3 shows the experimental results on simulated data and real data, respectively. Section 4 concludes the paper.

2 Methods

2.1 Basic terminology

In this paper, we only focus on the two-class problems, which can be simply extended to multiple classes as described in [22]. And we call one class as the *foreground* data set and the other as the *background* data set. Let D be a data set with two class labels, denoted by D=$\{D_1, D_2\}$, where D_1 corresponds to the *foreground* and D_2 to the *background*, and the number of instances is equal or not between them in different situations. The set of items that appear in all the instances is denoted as I = $\{i_1, i_2, i_3, \ldots, i_n\}$.

For a pattern p=$\{x_1, x_2, \ldots, x_r\}$, where $x_1, x_2, \ldots, x_{r-1}$ and x_r come from the set I, its occurrence in class D_t is denoted as Occ_t (t=1,2), and that in the whole data set is $Occ = Occ_1 + Occ_2$. We use *Sup(p, D)* to denote the support of pattern p in the data set D. This pattern is frequent in class D_t if its corresponding support $Sup(p, D_t) = \frac{Occ_t}{|D_t|}$ is no less than the user-specified threshold θ_{sup}. In fact, the discriminative pattern discovery can be interpreted as finding a set of patterns that are "over-expressed" in the *foreground* data set against the *background* data set. So we often only use the *foreground* to assess the frequency of a pattern.

Definition 1. p is a frequent pattern if *Sup(p, D_1)* is no less than θ_{sup}.

The set of candidate frequent patterns of size k is denoted as S_k. To discover all frequent patterns, the level-wise search strategy rooted from the Apriori algorithm is widely used in practice. It first accumulates the count for each pattern and collects those patterns with larger support than the given threshold θ_{sup} to form the set of frequent 1-patterns F_1. Subsequently, since a k-pattern will not be frequent if one of its sub-patterns of size k-1 is infrequent, F_{k-1} is utilized to generate S_k in an iterative way and those infrequent ones are pruned to generate F_k.

Owing to the fact that p consists of r items, we call p a r-pattern. This pattern has r sub-patterns of size r-1. These sub-patterns are denoted as p_1=$\{x_2, x_3, \ldots, x_r\}$, p_2=$\{x_1, x_3, \ldots, x_r\}$,$\ldots$, p_r=$\{x_1, x_2, \ldots, x_{r-1}\}$. We use p_j^i to denote the ith item of p_j (j=1,2,$\ldots$,r). And we describe the sets of instances where these sub-patterns occur as $D(p_1)$, $D(p_2)$,$\ldots$, $D(p_r)$, respectively. We utilize *Sig(p, D)* to denote the statistical significance calculation function for each pattern p, and it measures the discriminative power of p. In fact, the significance measurements that *DiffSup*, *odds ratio* and *relative risk* are widely used for evaluating discriminative patterns in different applications. More precisely, *DiffSup* is defined as the absolute difference of the relative supports of p in D_1 and D_2, and the corresponding significance value is calculated as:

$$Sig(p, D) = |Sup(p, D_1) - Sup(p, D_2)|. \tag{1}$$

That is, the patterns have the same chances to occur in the two classes results in *DiffSup* equals to 0. Thus, p can be considered as a valid discriminative pattern if its *DiffSup* greater than 0. Besides, both *odds ratio* and *relative risk* describe a likelihood change of the occurrence of one pattern between two classes. *Relative risk* is defined as the ratio of the supports of p in the two classes. When using *relative risk* to measure the discriminative power of p, the significance is:

$$Sig(p, D) = \frac{Sup(p, D_1)}{Sup(p, D_2)}. \tag{2}$$

A relative risk of 1 indicates that the target pattern under study is equally likely to occur in both classes. A relative risk greater than 1 means that this pattern is more likely to occur in the first class. In addition, the *odds* is the ratio of the probability that the interesting event does happen to the probability that it does not happen. The *odds ratio* is defined as the ratio of the odds of an event occurring in one class to the odds of it occurring in another class. If we adopt *odds ratio* to measure the discriminative power, then the significance of p is:

$$Sig(p, D) = \frac{Sup(p, D_1)(1 - Sup(p, D_1))}{Sup(p, D_2)(1 - Sup(p, D_2))} \tag{3}$$

Odds ratio has the same characteristics to *relative risk*: only those patterns whose *odds ratio* greater than 1 have potential to be the significant ones.

Particularly, if we consider $D(p_1)$, $D(p_2)$,..., $D(p_r)$ as the new data set instead of D when estimating the discriminative power, there are exactly r different significance values for p, that are $Sig(p, D(p_1))$, $Sig(p, D(p_2))$,..., $Sig(p, D(p_r))$, respectively. If we use above measures to assess the discriminative power, *Sig(p, D)* has positive correlation with the discriminative power: the bigger *Sig(p, D)* is, the more significant the pattern p is. For this reason, we use the minimum value of $Sig(p, D(p_1))$, $Sig(p, D(p_2))$,..., $Sig(p, D(p_r))$ as the *local* or *conditional* statistical significance in the assessment of discriminative power of each pattern. Then, the problem of conditional discriminative pattern mining is to discover all frequent patterns from D with sub-pattern derived statistical significance values pass the given threshold value.

Definition 2. Local statistical significance: $Sig_l(p, D) = \min(Sig(p, D(p_1)), Sig(p, D(p_2)),..., Sig(p, D(p_r)))$.

Overall, to identify all statistically significant, sufficiently frequent conditional discriminative patterns, we assess at least two aspects of each pattern: frequency and local statistical significance.

- **Frequency:** We impose the *support* constraint to reduce the search space and and prevent the generation of random artifacts.
- **Local statistical significance:** Note that the statistical evaluation of discriminative power for a pattern can be done in various ways. The statistical significance measures such as *DiffSup*, *relative risk* and *odds ratio* are available to be utilized interchangeably. The choice of significance assessment measure will not change the performance of our algorithms.

2.2 Problem formulation

As shown above, we strengthen and optimize the definition of discriminative pattern mining and try to conduct a non-redundancy discovery elegantly. The conditional discriminative patterns are deemed to be true positives with high discriminative power under no subsets interplays. However, there are two critical issues in the discriminative pattern finding. The first is how to ensure the conditional discriminative patterns are also significant under the traditional definition? The second is whether the effects of the sub-patterns can be removed through the use of local statistical significance?

We evaluate the statistical significance of a conditional discriminative pattern with the sets of instances induced from its sub-patterns. In contrast, the statistical significance of a pattern is estimated in the whole original data sets based on the traditional definition. The different calculation methods usually result in different values for discriminative power and there is no incidence relation with each other. Thus, we cannot guarantee these interesting patterns are also significant in the original data sets. For instance, consider the Figure 1 in the first section, which displays a sample dataset containing one conditional discriminative pattern. We have provided a detailed analysis with *DiffSup* to calculate the local significance of each pattern, and showed that p_1 is significant with the local significance 1.0. On the contratry, the statistical significance of p_1 is 0.6 indeed according to the conventional calculation method. Obviously, only when the threshold is set to be less than 0.6 we can ensure the conditional discriminative patterns are also true positive in the whole data sets. If we set the significance threshold to be 0.7, then p_1 is absolutely statistically significant under our definition while not based on the traditional definition. This illustrates that the conditional discriminative patterns may not be significant with the whole data sets as the background.

To address this issue, we impose an another significance constraint called *global* significance on each candidate pattern. We estimate the global significance $Sig_g(p, D)$ by considering the whole data set as the background according to the traditional definition. Besides, we call the statistical significance in accordance with our definition as the *local* significance. Hence, there are

two parameters to measure the significance of patterns: the global significance threshold θ_{g_sig} and the local significance threshold θ_{l_sig}, respectively. That is, to ensure the significance of one pattern p over D, two criteria must be satisfied simultaneously: $Sig_{g\_sig} \geq \theta_{g\_sig}$ and $Sig_{l\_sig} \geq \theta_{l\_sig}$.

To check if the use of conditional significance can remove the effect of sub-patterns in the assessment of discriminative power, we employ a measure called *improvement* proposed in [23] for justification. More precisely, the improvement is defined as the difference between the statistical significance of one pattern and that of its sub-patterns. In general, the positive improvement indicates that the discriminative power of target pattern comes from the combinations of all its constituent items rather than just one of its subsets. We should prune those redundant patterns that have non-positive improvements: the discriminative power of a pattern is equal to or less than its sub-patterns. To further clarify this issue, since we utilize the local significance to get rid of sub-pattern interplays, we calculate the difference values between the conditional discriminative k-pattern and its corresponding sub-patterns of k-1 with respect to the local significance.

Lemma 1. Each conditional discriminative pattern possesses positive improvement if *relative risk* is used to measure the discriminative power.

Proof. For a pattern p=$\{x_1, x_2, \ldots, x_r\}$ in data set D=$\{D_1, D_2\}$, suppose one of its sub-patterns p_j of length r-1 has the minimal significance value, i.e. $Sig_l(p, D) = Sig(p, D(p_j))$. We set $\theta_{l\_sig} = t$, $t \geq 1$. Then the Lemma 1 can be formulated as: if $Sig_l(p, D) \geq t$ then $Sig_g(p, D) > Sig_g(p_i, D)$ for any $1 \leq i \leq r$. □

$$Sig_l(p_i, D) \geq Sig_l(p, D) = Sig(p, D(p_j)) = \frac{Sup(p, D_1(p_j))}{Sup(p, D_2(p_j))} = \frac{|D_1(p)||D_2(p_j)|}{|D_2(p)||D_1(p_j)|} \geq t \tag{4}$$

$$Sig_g(p, D) = \frac{Sup(p, D_1)}{Sup(p, D_2)} = \frac{|D_2||D_1(p)|}{|D_1||D_2(p)|} \tag{5}$$

$$Sig_g(p_i, D) = \frac{Sup(p_i, D_1)}{Sup(p_i, D_2)} = \frac{|D_2||D_1(p_i)|}{|D_1||D_2(p_i)|} \tag{6}$$

Since t is a number no less than 1, so $\frac{|D_1(p)|}{|D_2(p)|} \geq \frac{|D_1(p_j)|}{|D_2(p_j)|}$. As a result, we can get that $Sig_g(p, D) > Sig_g(p_i, D)$ by making equation (5) minus equation (6), which returns a result greater than 0.

To make the following description easier to follow, we provide a precise problem definition of conditional discriminative pattern mining with clearly stated input and output.

- **Input:** A data set consists of two classes: the foreground data set D_1 and the background data set D_2, the support threshold θ_{sup}, the local significance threshold θ_{l_sig} and the global significance threshold θ_{g_sig}.
- **Output:** A set of conditional discriminative patterns, where each pattern $p \in R$ satisfies:(1) $Sup(p, D_1) \geq \theta_{sup}$; (2) $Sig_l(p, D) \geq \theta_{l\_sig}$; (3)$Sig_g(p, D) \geq \theta_{g\_sig}$.

2.3 Categorization of existing methods under the new formulation

Notice that Motif-X and MMFPh also measure the discriminative power of one pattern in a way that is similar to our definition of local significance. More precisely, although only θ_{l_sig} is used in both Motif-X and MMFPh, they implicitly require that one sub-pattern of the target pattern should be conditional significant as well. In this section, we first show that all the patterns reported by Motif-X and MMFPh are also global significant with high discriminative power as shown in the lemma below.

Lemma 2. For a pattern p=$\{x_1, x_2, \ldots, x_r\}$ in data set D=$\{D_1, D_2\}$, suppose we use relative risk as the measure and set both θ_{l_sig} and θ_{g_sig} to be t, $t \geq 1$. Let $p^{(k)} = \{x_i \mid i \leq k\}$. If $Sig(p^{(k)}, D(p^{(k-1)})) \geq t$, for all $1 \leq k \leq r$, then $Sig_g(p, D) \geq t$.

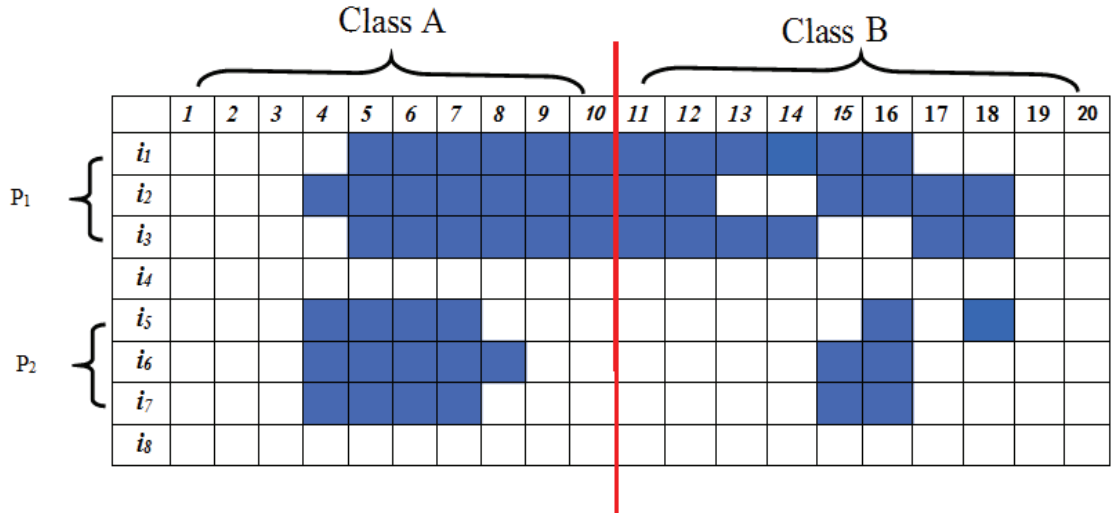

Figure 2: An sample data set containing twenty instances divided into two classes, eight items making up two different categorical patterns. Lines mean items, and columns indicate instances. Cells in shade represent current items exist in the corresponding instances and vice versa. For example, cell (i_1, 5) is in shade, so the fifth instance in Class A contains item i_1.

Proof. □

$$Sig(p^{(k)}, D(p^{(k-1)})) = \frac{Sup(p^{(k)}, D_1(p^{(k-1)}))}{Sup(p^{(k)}, D_2(^{(k-1)}))} = \frac{|D_1(p^{(k)})||D_2(p^{(k-1)})|}{|D_2(p^{(k)})||D_1(p^{(k-1)})|} \tag{7}$$

$$Sig(p^{(k-1)}, D(p^{(k-2)})) = \frac{Sup(p^{(k-1)}, D_1(p^{(k-2)}))}{Sup(p^{(k-1)}, D_2(p^{(k-2)}))} = \frac{|D_1(p^{(k-1)})||D_2(p^{(k-2)})|}{|D_1(p^{(k-2)})||D_2(p^{(k-1)})|} \tag{8}$$

$$Sig_g(p, D) = \frac{Sup(p, D_1)}{Sup(p, D_2)} = \frac{|D_1(p)||D_2|}{|D_1||D_2(p)|} \tag{9}$$

It is easy to see that the equation $Sig(p^{(k)}, D(p^{(k-2)})) \geq t^2$ by multiplying equation (7) and (8). That is, $p^{(k)}$ is statistically significant in the data sets derived from $p^{(k-2)}$. This rule also applies to pattern $p^{(k-2)}$ so that $p^{(k-2)}$ is significant in the instances induced by its one sub-patterns as well. Therefore, we can infer that $Sig_g(p, D) \geq t^r$ by iterating the multiplication process. Since t is a positive number that is larger than 1, then $Sig_g(p, D)$ must be greater than t, too. Thus, p is global significant without doubt.

Lemma 2 shows that Motif-X and MMFPh can find patterns that are both locally and globally significant, however, they may miss some meaningful patterns and cannot guarantee the completeness. Figure 2 provides such an example.

Suppose Class A corresponds to the foreground data set and Class B corresponds to the background data set. We set the significance threshold $\theta_{sig} = 0.3$ and the support threshold $\theta_{sup} = 0.2$. For illustration purpose, we adopt *DiffSup* as the significance measure here. Under this setting, P_1 and P_2 are definitely significant patterns according to the discriminative pattern definition. However, the sub-patterns of P_1, that are $\{i_1, i_2\}$, $\{i_1, i_3\}$ and $\{i_2, i_3\}$, are all insignificant since *DiffSup*=0.2 is less than θ_{sig}. In contrast, all the sub-patterns of size 2 of P_2 are statistically significant with *DiffSup* equals to 0.3. Thus, methods like Motif-X and MMFPh must filter out P_1 and only report P_2 because there is no chance to generate and evaluate P_1 since it lacks of frequent and significant constituent patterns. Clearly, they may miss some statistically significant patterns like P_1 and provide no guarantee on the completeness of reported pattern set.

In summary, the algorithms that only adopt global significance such as Motif-All often return a set of discriminative patterns that contains lots of false positives whose discriminative power

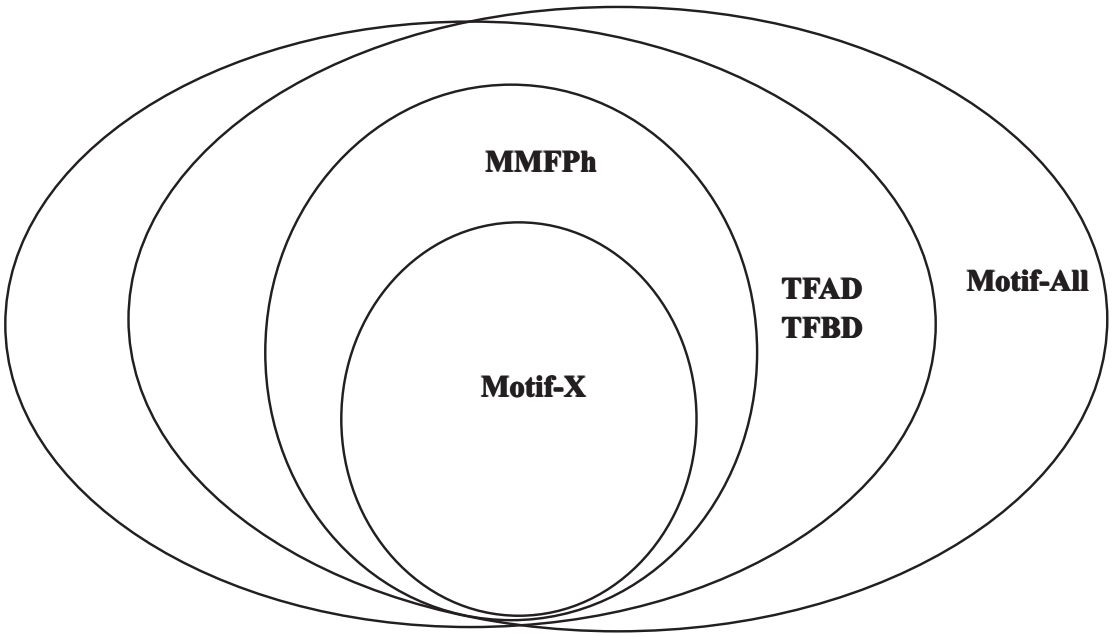

Figure 3: The relationship of the sets of significant patterns discovered by different algorithms.

mainly benefits from the subsets. Furthermore, it has been shown that MMFPh is able to find much more discriminative patterns missed by Motif-X. While our formulation and proposed algorithms (TFAD, TFBD) use the global significance together with the local significance so as to report a set of patterns that are both statistically significant and non-redundant. The detailed illustration of the relationship of the pattern sets reported by the different algorithms is provided in Figure 3.

2.4 TFAD algorithm

The TFAD algorithm (Algorithm 1) tests the frequency together with the discriminative power of a pattern at the same time. For instance, when discovering the set of F_1 from S_1 in the first iteration, we also calculate the local significance for each frequent member of F_1 in the meanwhile. The global significance of 1-patterns is identical to their local significance. Since one pattern is statistically significant on condition that its global significance and local significance are not less than the given threshold θ_{g_sig} and θ_{l_sig}, respectively. We filter out the insignificant patterns with $Sig_g(p, D) \geq \theta_{g\_sig}$ and $Sig_l(p, D) \geq \theta_{l\_sig}$, and then add the rest to the result set. So after investigating every possible 1-pattern, all the patterns in the result set are significantly conditional discriminative patterns of size 1. With respect to the kth iterations, we perform the following operations:

1. Generate the set of potential frequent patterns of size k, i.e. S_k, based on the frequent patterns set F_{k-1}. For all patterns in S_k, we first calculate their supports. If one k-pattern is infrequent, then we prune it immediately. Whereas if one k-pattern is frequent, then we add it to F_k first and then we use its (k-1)-sub-patterns p_h to generate the corresponding data sets that contain the current sub-patterns, denoted by $D(p_h)$. We replace the original data sets by the new data sets $D(p_h)$ to calculate the local significance value for the target k-pattern. Since one k-pattern owes k (k-1)-sub-patterns, we choose the minimum $Sig(p, D(p_h))$ to assess the discriminative power of a pattern and filter out insignificant patterns that may correspond to random artifacts. Then we calculate the global significance of the target pattern with the original data sets. Those patterns whose global and local significance are greater than θ_{g_sig} and θ_{l_sig} correspondingly are added to R.

Algorithm 1 $TFAD(D, \theta_{sup}, \theta_{l\_sig}, \theta_{g\_sig})$

Input: Dataset D with class labels i, denoted by $D_i (i = 1, 2)$,
thresholds θ_{sup}, θ_{l_sig} and θ_{g_sig}
Output: The set of conditional discriminative patterns R
$R \leftarrow \emptyset$
$F_1 \leftarrow \{p \in S_1 \mid Sup(p, D_1) \geq \theta_{sup}\}$ //frequent pattern of size 1
$R \leftarrow R \cup \{p \in F_1 \mid Sig(p, D) \geq \theta_{sig}\}$//discriminative patterns
$k \leftarrow 1$
while $F_k \neq \emptyset$ **do**
  $k \leftarrow k + 1$
  $S_k \leftarrow F_{k-1} \bowtie F_{k-1}$
  for each $p \in S_k$ **do**
    if $Sup(p, D_1) \geq \theta_{sup}$ **then**
      add p to F_k
      for $h = 1$ to k **do**
        $p_h \leftarrow \{p_h^i \mid i \neq h\}$
        $D(p_h) \leftarrow \{ins \in D \mid \forall_j : ins_j = p_h^j\}$
        $h\_Sig \leftarrow Sig(p, D(p_h))$// local significance for p by using $D(p_h)$ as the new data
      end for
      $Sig_l(p, D) \leftarrow min(h\_Sig)$
      $Sig_g(p, D) \leftarrow Sig(p, D)$
      if $Sig_l(p, D) \geq \theta_{l\_sig}$ and $Sig_g(p, D) \geq \theta_{g\_sig}$ **then**
        $R = R \cup \{p\}$// conditional discriminative patterns
      end if
    end if
  end for
end while

2. We repeat this process until no more frequent patterns can be discovered in S_k. We return R as the final result.

2.5 TFBD algorithm

TFBD tests the frequency before calculating the discriminative power of each pattern. The detailed pseudocode is provided in Algorithm 2. TFBD divides the mining process into two stages. This method performs frequent pattern mining first, and later verify the characteristic with respect to discriminative power of those candidate patterns produced in first stage.

1. Perform the frequent pattern mining. Generate the set of potential frequent patterns of size k, denoted by S_k, based on the frequent patterns set F_{k-1}. Prune less frequent patterns from S_k to form F_k. Repeat a number of rounds until no frequent patterns can be found. Therefore the result set $F = \cup F_k$ in this step is a collection of all frequent patterns.

2. Test both global and local statistical significance of each frequent pattern. We test the statistical significance of each pattern as done in Algorithm 1 until all the candidate patterns produced in the first step have been examined.

Algorithm 2 $TFBD(D, \theta_{sup}, \theta_{l\_sig}, \theta_{g\_sig})$

Input: Dataset D with class labels i, denoted by $D_i (i = 1, 2)$,
thresholds θ_{sup}, θ_{l_sig} and θ_{g_sig}
Output: The set of conditional discriminative patterns R
$R \leftarrow \emptyset$
$F_1 \leftarrow \{p \in S_1 \mid Sup(p, D_1) \geq \theta_{sup}\}$ //frequent pattern of size 1
$k \leftarrow 1$
while $F_k \neq \emptyset$ **do**
  $k \leftarrow k + 1$
  $S_k \leftarrow F_{k-1} \bowtie F_{k-1}$
  for each $p \in S_k$ **do**
    if $Sup(p, D_1) \geq \theta_{sup}$ **then**
      add p to F_k
    end if
  end for
end while
$F = \cup F_k$
for all patterns $p \in F$ **do**
  for $h = 1$ to k **do**
    $p_h \leftarrow \{p_h^i \mid i \neq h\}$
    $D(p_h) \leftarrow \{ins \in D \mid \forall_j : ins_j = p_h^j\}$
    $h\_Sig \leftarrow Sig(p, D(p_h))$// significance for p by using $D(p_h)$ as the new data
  end for
  $Sig_l(p, D) \leftarrow min(h\_Sig)$
  $Sig_g(p, D) \leftarrow Sig(p, D)$
  if $Sig_l(p, D) \geq \theta_{l\_sig}$ and $Sig_g(p, D) \geq \theta_{g\_sig}$ **then**
    $R \leftarrow R \cup p$ // conditional discriminative patterns
  end if
end for

2.6 Proofs of completeness and correctness

Theorem 1. Both TFAD and TFBD algorithms are complete.

Proof. The completeness of TFAD and TFBD can be shown by the following two facts. The first is that the Apriori algorithm is complete; that is, it explicitly enumerates and checks all frequent patterns in the mining process. The second is that these two methods only prune those insignificant patterns with respect to both global significance and local significance. □

Theorem 2. Both TFAD and TFBD algorithms are correct.

Proof. TFAD and TFBD assess two aspects of each pattern: frequency and statistical significance including global significance and local significance. They identify all true discriminative patterns. This is due to the fact that they evaluate them by the subsets of the original data sets induced by their sub-patterns instead. One pattern is considered to be potential significant if its local significance value satisfies greater than θ_{l_sig}. It is this strategy that enables us to get rid of sub-pattern interactions and remove the false positives whose discriminative power induced from sub-patterns. Besides, all the patterns must be also significant according to the traditional definition on condition that their global significance values are greater than θ_{g_sig}. □

3 Experimental Results

In order to demonstrate the efficacy and utility of our algorithms, we conduct a series of tests with real data. In our experiments, we compare our algorithms against the Motif-All algorithm and the MMFPh algorithm with respect to efficiency, completeness and accuracy. Note that several methods have been proposed to detect discriminative patterns, the reason why we choose Motif-All and MMFPh as comparisons here is that they are representatives for algorithms that use only global significance threshold and local significance threshold, respectively. Besides, we use the same discriminative measure in all algorithms so as to make their outputs comparable. More precisely, we choose odds ratio since a p-value can be derived. That is, we first adopt large sample approximations to the sampling distribution of the log odds ratio for statistical inference, and then we can get z-value by making log odds ratio divided by its standard error. The z-value follows a standard normal distributions so that we can calculate the p-value to assess the statistical significance of each pattern.

In the experiments, we apply TFAD, TFBD, MMFPh and Motif-All to real data including phosphorylation data sets, a breast cancer gene expression data set and SNP data. The details of these data sets are provided in the following sections. In each experiment using one kind of data, we first present a brief description of the data and tune the thresholds so as to clarify the comparison. Subsequently, we perform a general analysis of the patterns discovered and then illustrating the superiority of TFAD and TFBD against MMFPh and Motif-All.

3.1 Phosphorylation data

3.1.1 Data description

Phosphorylation data set [8] is composed of ten groups of phosphorylated peptides set served as the *foreground* and unphosphorylated peptides set served as the *background*. Protein phosphorylation is an essential post-translational modification event for the regulation and maintenance for most biological processes. A phosphorylated peptide is defined as one peptide with at least one residue, which is denoted with a underlined character (S, T or Y), known to be phosphorylated. If the residue can not be phosphorylated, we call that peptide as unphosphorylated peptide. Here all the peptides have the fixed length 13 and they are aligned on the residue that lies in the center position. Each data set consists of about 5000 phosphorylated peptides or 5000 unphosphorylated peptides as instances. Particularly, the same phosphorylated residue is not counted as items. For example, PSxD is a 2-pattern with two items, P and D, and 'x' represents that position can be replaced by any arbitrary items.

3.1.2 Results

To make the comparison more remarkable, we choose a lower value as the support threshold and a higher value as the significance threshold so as to discover more patterns. For consistency, we take the support threshold θ_{sup} as 0.01 and the p-value threshold θ_{sig} as 0.1 for all the algorithms under discussion. The questions we want to answer in this experiment are: How many discriminative patterns of different sizes can be discovered by the four algorithms, respectively? Which methods can detect more conditional discriminative patterns?

Table 1 summarizes the details of discriminative patterns discovery by TFAD, TFBD, MMFPh and Motif-All. Several observations can be made from Table 1.

First, TFAD and TFBD are able to find more statistically significant patterns than MMFPh in general. Besides, Motif-All report much more patterns, and all the reported patterns of TFAD, TFBD and MMFPh are included in the result set of Motif-All.

Second, we find that all the algorithms present a same set of interesting 1-patterns. Besides, MMFPh also returns the same set of 2-patterns as TFAD and TFBD while Motif-All discovers much more by including many other 2-patterns which are pruned in the former three methods. Careful detailed analysis of these 1-patterns and 2-patterns shows that those reported by TFAD, TFBD and MMFPh are all conditional discriminative patterns according to our definition. In contrast, some of the rest patterns only presented by Motif-All turn out to be false positives whose discriminative power mainly benefits from their subsets. For example,LxxxxxSP, a pattern of size 2, is composed of LxxxxxS and SP. We find that SP is one member of the set of reported 1-patterns while LxxxxxS is not. This means the former is statistical significant with high discriminative power, whereas the latter is not. The discriminative part conceals the performance of the nondiscriminative part and promotes the power of the whole body. So we regard patterns like this as false positives as demonstrated in section 1. In addition, we will not report some other patterns only discovered by Motif-All that each part has similar high discriminative power to the whole pattern. SxxxxxSP as an example, both SxxxxxS and SP are all statistical significant so that SxxxxxSP has little improvement of discriminative power compared to that of its subsets.

Third, there is visible difference in the discovery of discriminative 3-patterns. MMFPh fails to find any 3-patterns while Motif-All report five 3-patterns that are two more than TFAD and TFBD. All the reported 3-patterns are listed in Table 2. We first re-do significance test according to our definition on the intersection of the resulting 3-pattern set of TFAD, TFBD and Motif-All. They are definitely true positive with accurate significance. Subsequently, we apply our significance test to the two patterns, SPxSP and PxSPxS, only discovered by Motif-All. SPxSP is composed of sub-patterns that 1-pattern SxxS and 2-pattern PxSP. SxxS is not a discriminative pattern while PxSP is a discriminative pattern with p-value 0.04 smaller than $\theta_{sig} = 0.1$. Thus, SPxSP is a nondiscriminative pattern according to our definition although it has a bigger significance value than θ_{sig}. Similar analysis can also be made for PxSPxS. Hence, SPxSP and PxSPxS are false positives and filtering out the patterns like this is just our target in the discriminative pattern discovery as discussed in section 1. The largest size of interesting discriminative patterns is 3 under the setting of $\theta_{sup} = 0.01$ and $\theta_{sig} = 0.1$.

Since there is little difference in performance of TFAD, TFBD and MMFPh and even Motif-All when the target patterns of size smaller than 3. Obviously, the smaller size is, the more similar the result set of the algorithms is. This is clearly visible in the result on phosphorylation data as showed in Table 1. Thus, we can infer the increase of size of target patterns will apparently highlight the advantage of TFAD and TFBD over MMFPh and Motif-All.

To further check if this is true, we also perform discriminative pattern mining at the support threshold of 0.1 with unchanged significance threshold. We obtain an extremely identical set of fourteen interesting patterns by these four methods under this setting. When change the support threshold to 0.005, TFAD and TFBD report 843 significant patterns, whereas MMFPh report 813 patterns and Motif-All report 2300 patterns. Accordingly, MMFPh misses a number of interesting patterns while Motif-All reports many false positives. This demonstrates that TFAD and TFBD

not only can find more significant patterns than MMFPh but also are more qualified than Motif-All in a flexible manner.

We design the same tests on the other groups of phosphorylation data sets, and change the threshold values for the algorithms by degrees, they show similar performance in the discriminative pattern discovery. In conclusion, Table 1 illustrates that MMFPh is useful and effective in presenting true positives, whereas it would miss some statistically significant patterns like as showed in Table 2. On the other hand, Motif-All achieve completeness at the cost of involving lots of false positives whose discriminative power induced from sub-patterns. In contrast, TFAD and TFBD can not only find all the true positives but also without any false positives. Hence, the empirical comparison shows that TFAD and TFBD outperform the other methods like MMFPh and Motif-All, discovering all the discriminative patterns with real statistical significance.

Table 1: The number of reported patterns of different size on one group of phosphorylation data sets. Here the support threshold θ_{sup} is 0.01 and the significance threshold θ_{sig} is 0.1.

Method	Size				Total
	1	2	3	Others	
TFAD	57	275	3	0	335
TFBD	57	275	3	0	335
MMFPh	57	275	0	0	332
Motif-All	57	615	5	0	687

Table 2: Patterns of size 3 reported by TFAD, TFBD, MMFPh and Motif-All on one group of phosphorylation data sets.

Method	3-patterns	Total
TFAD	SPxx<u>S</u>P PxPx<u>S</u>P <u>S</u>PxxSP	3
TFBD	SPxx<u>S</u>P PxPx<u>S</u>P <u>S</u>PxxSP	3
MMFPh		0
Motif-All	SPxx<u>S</u>P SPx<u>S</u>P PxPx<u>S</u>P Px<u>S</u>PxS <u>S</u>PxxSP	5

3.2 Breast cancer gene expression data

3.2.1 Data description

The breast cancer gene expression data set derived from [6] is a binary data set that has 11962 items and 295 transactions. It is constructed from a complex, real data with the expression profiles of 25,000 genes in 295 breast cancer patients. We categorize the patients into two classes with respect to whether the patient survives the disease or not, where the former corresponds to the

foreground and the latter corresponds to the *background*. To date, we only focus on 5,981 genes as in [6] since their occurrence is significantly different with at least a twofold change between the two classes and their expression measurements are accurate (p-value ≤ 0.01) for at least five patients. The genes under discussion are considered to make more sense in practice. Between-gene variations have been properly eliminated to normalize the data. This data set is stored in a binary table. Two binary columns together present the information of a single gene: a 1 in the first column means the expression of the gene is less than -0.2, whereas a 1 in the second column indicates the expression of the gene is greater than 0.2. The genes whose expression values between -0.2 and 0.2 are not included in this data set since they are expected to be uninteresting by involving substantial noise.

Discriminative pattern mining on this data set can facilitate us to find out the pathogeny of breast cancer and even its cure. Our experiments are designed to evaluate the efficacy of different algorithms and analyze their performance for this task.

3.2.2 Results

For this data set, we first set the support threshold $\theta_{sup} = 0.6$ and the significance threshold $\theta_{sig} = 10^{-6}$ randomly. See the Table 3 for the details of the significant patterns found by TFAD, TFBD, Motif-All and MMFPh respectively under this setting. Accordingly, MMFPh reports the minimum number of patterns, whereas Motif-All produces the most that beyond enough. TFAD and TFBD presents an identical set of discriminative patterns that less than Motif-All while more than MMFPh.

For the confirmation the performance of TFAD and TFBD, we first investigate each reported pattern by the two methods, and then analyze their characteristics with respect to the discriminative power. We also examine the significance of every sub-pattern constituting these patterns. Facts shows that each pattern is no doubt statistically significant with accurate significance. In addition, every part not only contributes equally but also is inferior to their whole combination about the capability of the discriminative power. Along with we utilize enough rigorous definition of discriminative patterns, and conduct the discriminative pattern mining in a strict way and have re-tested their significance, so all the patterns must be true positives without no sub-pattern interplays. Considering TFAD and TFBD enumerate and check all the potential candidate patterns, both TFAD and TFBD present a complete set of interesting patterns delicately.

Indeed, Motif-All also succeeds in finding all the significant discriminative pattern reported by TFAD and TFBD, whereas including some other controversial ones that need to be further discussed. For this reason, we have to re-conduct significance tests for those patterns only found by Motif-All. We first extract them to form a new set, and then check the statistical significance of each pattern together with that of its every component part. All of these patterns have something in common that they consist of significant parts and insignificant parts, and their significant parts enormously contribute to the discriminative power of the whole patterns. According to our definition, all of these patterns are considered as false positives with negative statistical significance. Therefore, due to the weaker pruning of the global significance in the assessment of discriminative power, Motif-All often fails to get rid of subsets interplays and returns a larger number of patterns with lots of undesired noise.

Specially, we find that all the patterns found by MMFPh also occur in the result sets reported by the other algorithms. What's more, it has been shown that all the patterns reported by TFAD and TFBD are true positives, this illustrates MMFPh can ensure the valid significance of each reported pattern as well. However, on one hand, the number of patterns found by MMFPh is much less than TFAD and TFBD; on the other hand, both TFAD and TFBD can guarantee the correctness and effectivity of all their patterns. Hence, MMFPh succeeds in finding true positives while lacking completeness. For instance, {19, 331, 3387} is a pattern reported by TFAD, TFBD and Motif-All but missed by MMFPh, where the number stands for the gene in that column in the binary table. More precisely, 19 indicates the gene Contig56678_RC, 331 corresponds to the gene

ARF6 and 3387 means the gene AL133619. According to the discriminative pattern definition, this pattern is indeed a significant pattern with high discriminative power. In contrast, its sub-patterns {19, 311}, {311, 3387} and {19, 3387} are all insignificant ones with lower discriminative power. As shown in the former section, since there is no frequent and significant sub-pattern for {19, 331, 3387}, so there is no chance to generate and evaluate the target pattern in MMFPh process, and thus MMFPh considers it as negative and prunes it for certain. In this sense, MMFPh fails to discover some interesting patterns because of its incomplete search. In addition, we also find that most of those meaningful patterns missed by MMFPh are of larger sizes. Therefore, MMFPh is able to remove the sub-pattern interactions in the discriminative pattern discovery at the cost of completeness.

For illustration purpose, we also study the performance of the different algorithms under variation in the statistical significance standard and support level. We gradually set a relatively bigger significance threshold and a relatively smaller support threshold to obtain more patterns. All the methods finish their mining within an acceptable amount of time. The results are summarized and provided in the Table 4. Under any setting, it is easy to see Motif-All always reports the most of patterns and MMFPh returns the least. If we adopte variable-controlling approach by lowering the support threshold with unchanged significance threshold, or increasing the significance threshold with unchanged support threshold, there will be much more interesting patterns of different sizes that are reported. Consequently, we can infer that the larger the support threshold and the smaller the significance threshold are, the more significant patterns of the larger sizes are found, and the more obvious difference between the results of the algorithms. This is because if one pattern is of a certain especially larger size, then this pattern contains more sub-patterns, and the interplays play more important role and are more potential to effect the whole combination with respect to discriminative power, especially compared to one pattern of a little smaller size. Hence, it increases the probability for Motif-All consider the target pattern as an interesting one even its discriminative power induces from some subsets, or for MMFPh filters it out if it is a positive one but with no frequent and significant sub-patterns. In this sense, Motif-All and MMFPh may be more appropriate for discovering those patterns of smaller size. To conclude, TFAD and TFBD are better algorithms with both completeness and correctness in discriminative pattern discovery.

Table 3: The number of reported patterns of different size on the breast cancer gene expression data sets. Here the support threshold θ_{sup} is 0.6 and the significance threshold θ_{sig} is 10^{-6}.

Method	Size		Total of patterns
	Number of size	The largest size	
TFAD	3	3	87
TFBD	3	3	87
MMFPh	3	3	62
Motif-All	4	4	276

Table 4: The number of reported patterns of different size on the breast cancer gene expression data sets by tuning the thresholds θ_{sup} and θ_{sig} gradually.

Threshold	Method	Size		Total of patterns
		Number of size	The largest size	
$\theta_{sup} = 0.5$ $\theta_{sig} = 10^{-6}$	TFAD	4	4	137
	TFBD	4	4	137
	MMFPh	4	4	114
	Motif-All	5	5	469
$\theta_{sup} = 0.3$ $\theta_{sig} = 10^{-6}$	TFAD	5	5	243
	TFBD	5	5	243
	MMFPh	5	5	198
	Motif-All	7	7	642
$\theta_{sup} = 0.3$ $\theta_{sig} = 10^{-3}$	TFAD	7	7	1915
	TFBD	7	7	1915
	MMFPh	6	6	1458
	Motif-All	8	7	7858

References

[1] Cheng, H., Yan, X., Han, J., and Hsu, C.-W. (2007). Discriminative Frequent Pattern Analysis for Effective Classification. In *Proceeding of the international Conference on Data Engineering*, pages 716–725.

[2] Cong, G., Tan, K.-L., and Anthony K.H.Tung, X. X. (2006). Mining Top-K Covering Rule Groups for Gene Expression Data. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 670–681.

[3] del Jesus, Jos, M., Gonzlez, Pablo, P., Guzmn, H. F., and Mesonero, M. (2007). Evolutionary fuzzy rule induction process for subgroup discovery: A case study in marketing. In *Proceedings of the sixth European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 578–592.

[4] Dong, G. and Li, J. (1999). Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 43–52.

[5] Fang, G., Wang, W., Oately, B., Ness, B. V., Steinbach, M., and Kumar, V. (2011). Characterizing Discriminative patterns. Arxiv preprint arXiv:1102.4104.

[6] Fang, G., Pandey, G., Wang, W., Gupta, M., Steinbach, M., and Kumar, V. (2012). Mining Low-support Discriminative Patterns from Dense and High-Dimensional Data. *IEEE Transactions on Knowledge and Data Engineering*, **24**(2), 279–294.

[7] Gong, H. and He, Z. (2012). Permutation methods for testing the significance of phosphorylation motifs. *Statistics and Its Interface*, **5**, 61–73.

[8] Gong, H., Liu, X., Wu, J., and He, Z. (2013). Data construction for phosphorylation site prediction. *Briefings in bioinformatics*.

[9] Good, P. I. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses*, chapter 9. Springer-Verlag New York Inc., New York.

[10] He, Z., Yang, C., Guo, G., Li, N., and Yu, W. (2011). Motif-All: Discovering All Phosphorylation Motifs. *BMC Bioinformatics*, **12:S22**.

[11] I.Webb, G., Butler, S., and Newlands, D. (2003). On detecting differences between groups. In *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 256–265.

[12] Klösgen, W. (1996). *Explora: a multipattern and multistrategy discovery assistant*, pages 249–271. American Association for Artificial Intelligence, Menlo Park, CA.

[13] Klösgen, W. and May, M. (2002). Spatial Subgroup Mining Integrated in an Object-Relational Spatial Database. In *Proceedings of the sixth European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 275–286.

[14] Kralj, P., Lavrač, N., and Dragan Gamberger, A. K. (2007). Contrast Set Mining for Distinguishing between Similar Diseases. In *Proceedings of the conference on Artificial Intelligence in Medicine*, pages 109–118.

[15] Li, J., Ramamohanarao, K., and Dong, G. (2000). The space of Jumping Emerging Patterns and Its Incremental Maintenance Algorithms. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 551–558.

[16] Li, W., Han, J., and Pei, J. (2001). CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. In *Proceedings of the 2001 IEEE Industrial Conference on Data Mining*, pages 369–376.

[17] Liu, B., Ma, Y., and Wong, C. K. (2001). Integrating Classification and Association Rule Mining. In *Proceedings of ACM SIGKDD Industrial Conference Knowledge Discovery on Data Mining*, pages 80–86.

[18] Ma, L., L.Assimes, T., B.Asadi, N., Iribarren, C., Quertermous, T., and H.Wong, W. (2010). An almost exhaustive search-based sequential permutation method for detecting epistasis in disease association studies. *Genetic Epidemiology*, **34**, 434–443.

[19] Novak, P. K. and Lavrač, N. (2009). Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast set, Emerging Pattern and Subgroup Mining. *Journal of Machine Learning Research*, **10**, 377–403.

[20] Novak, P. K., Lavrač, N., Gamberger, D., and Krstačić., A. (2009). CSM-SD: Methodology for Contrast Set Mining through Subgroup Discovery. *Journal of biomedical informatics*, **42**, 113–122.

[21] Schwart, D. and Gygi, S. P. (2005). An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nature Biotechnology*, **23**, 1391 – 1398.

[22] Stephen, D. B. and Michael, J. P. (1999). Detecting change in categorical data: Mining contrast sets. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 302–306.

[23] Stephen, D. B. and Michael, J. P. (2001). Detecting Group Differences: Mining Contrast Sets. *Data Mining and Knowledge Discovery*, **5**, 213–246.

[24] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*, chapter 6. Addison-Wesley, New Jersey.

[25] Wang, T., N.Kettenbach, A., A.Gerber, S., and Bailey-Kellogg, C. (2012). MMFPh: A Maximal Motif Finder for Phosphoproteomics Datasets. *Bioinformatics*, **28**, 1562–1570.

[26] Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *Proceedings of the 1st European Conference on Principles and Practice of Knowledge Discovery*, pages 78–87.

[27] Yin, X. and Han, J. (2003). CPAR: Classification Based on Predictive Association Rules. In *Proceedings of the Third SIAM International conference on Data Mining*, pages 331–335.