Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 09-011

Mining Low-Support Discriminative Patterns from Dense and High-Dimensional Data

Gang Fang, Gaurav Pandey, Wen Wang, Manish Gupta,
Michael Steinbach, and Vipin Kumar

April 02, 2009

Mining Low-Support Discriminative Patterns from Dense and High-dimensional Data

Gang Fang, Gaurav Pandey, Wen Wang, Manish Gupta, Michael Steinbach, *Member, IEEE,* Vipin Kumar, *Fellow, IEEE*

Gang Fang, Wen Wang, Michael Steinbach and Vipin Kumar are with the Department of Computer Science and Engineering, University of Minnesota, 200 Union Steet SE, Minneapolis, MN 55455. E-mail: $\{gangfang, wwang, steinbac, kumar\}$ $@cs.umn.edu$; Gaurav Pandey is with the Department of Plant and Microbial Biology, University of California, Berkeley. E-mail: $gaurav@compbio.berkeley.edu$; Manish Gupta is with Oracle India Private Ltd. E-mail: $manishgupta.iitg@gmail.com$

Abstract

Discriminative patterns can provide valuable insights into datasets with class labels, that may not be available from the individual features or the predictive models built using them. Most existing approaches work efficiently for sparse or low-dimensional datasets. However, for dense and high-dimensional datasets, they have to use high thresholds to produce the complete results within limited time, and thus, may miss interesting low-support patterns. In this paper, we address the necessity of trading off the completeness of discriminative pattern discovery with the efficient discovery of low-support discriminative patterns from such datasets. We propose a family of anti-monotonic measures named $SupMaxK$ that organize the set of discriminative patterns into nested layers of subsets, which are progressively more complete in their coverage, but require increasingly more computation. In particular, the member of $SupMaxK$ with $K = 2$, named $SupMaxPair$, is suitable for dense and high-dimensional datasets. Experiments on both synthetic datasets and a cancer gene expression dataset demonstrate that there are low-support patterns that can be discovered using $SupMaxPair$ but not by existing approaches. Furthermore, we show that the low-support discriminative patterns that are only discovered using $SupMaxPair$ from the cancer gene expression dataset are statistically significant and biologically relevant. This illustrates the complementarity of $SupMaxPair$ to existing approaches for discriminative pattern discovery. The codes and dataset for this paper are available at http://vk.cs.umn.edu/SMP/.

Index Terms

Association analysis, Discriminative pattern mining, Biomarker discovery, Permutation test

I. Introduction

For data sets with class labels, association patterns [2], [43] that occur with disproportionate frequency in some classes versus others can be of considerable value in many applications. Such applications include census data analysis that aims at identifying differences among demographic groups [14], [5] and biomarker discovery, which searches for groups of genes or related entities, that are associated with diseases [8], [39], [1]. We will refer to these patterns as discriminative patterns[1] in this paper, although they have also been investigated under other names [35], such as emerging patterns [14] and contrast sets [5]. In this paper, we focus on 2-class problems, which can be generalized to multi-class problems as described in [5].

Discriminative patterns have been shown to be useful for improving the classification performance for data sets where combinations of features have better discriminative power than

[1] The terms ”pattern” and ”itemset” are used interchangeably in this paper.

the individual features [9], [13], [47], [10], [15], [30]. More importantly, as discussed in [5], discriminative pattern mining can provide insights beyond classification models. For example, for biomarker discovery from case-control data (e.g. disease vs. normal samples), it is important to identify groups of biological entities, such as genes and single-nucleotide polymorphisms (SNPs), that are collectively associated with a certain disease or other phenotypes [1], [50], [38]. Algorithms that can discover a comprehensive set of discriminative patterns are especially useful for domains like biomarker discovery, and such algorithms are the focus of this paper.

The algorithms for finding discriminative patterns usually employ a measure for the discriminative power of a pattern. Such measures are generally defined as a function of the pattern's relative support[2] in the two classes, and can be defined either simply as the ratio [14] or difference [5] of the two supports, or other variations, such as its information gain [9], Gini index, odds ratio [43] etc. In this paper, we use the measure that is defined as the difference of the supports of an itemset in the two classes (originally proposed in [5] and used by its extensions [24], [25]). We will refer to this measure as $\mathit{DiffSup}$ (formally discussed in Section II). Given a dataset with 0-1 class labels and a $\mathit{DiffSup}$ threshold r, the patterns with $\mathit{DiffSup} \geq r$ can be considered as valid discriminative patterns.

To introduce some key ideas about discriminative patterns and make the following discussion easier to follow, consider Figure 1, which displays a sample dataset[3] containing 15 items (columns) and two classes, each with 10 instances (rows). In the figure, four patterns (sets of binary variables) can be observed: $P_1 = \{i_1, i_2, i_3\}$, $P_2 = \{i_5, i_6, i_7\}$, $P_3 = \{i_9, i_{10}\}$ and $P_4 = \{i_{12}, i_{13}, i_{14}\}$. P_1 and P_4 are interesting discriminative patterns that occur with different frequencies in the two classes, whose $\mathit{DiffSup}$ is 0.6 and 0.7 respectively. In contrast, P_2 and P_3 are uninteresting patterns with a relatively uniform occurrence across the classes, both having a $\mathit{DiffSup}$ of 0. Furthermore, P_4 is a discriminative pattern whose individual items are also highly discriminative, while those of P_1 are not. Based on support in the whole dataset, P_2 is a frequent non-discriminative pattern, while P_3 is a relatively infrequent non-discriminative pattern.

[2] Note that, in this paper, unless specified, the support of a pattern in a class is relative to the number of transactions (instances) in that class, i.e. a ratio between 0 and 1, which can help handle the case of skewed class distributions.

[3] The discussion in this paper assumes that the data is binary. Nominal categorical data can be converted to binary data without loss of information, while ordinal categorical data and continuous data can be binarized, although with some loss of magnitude and order information.

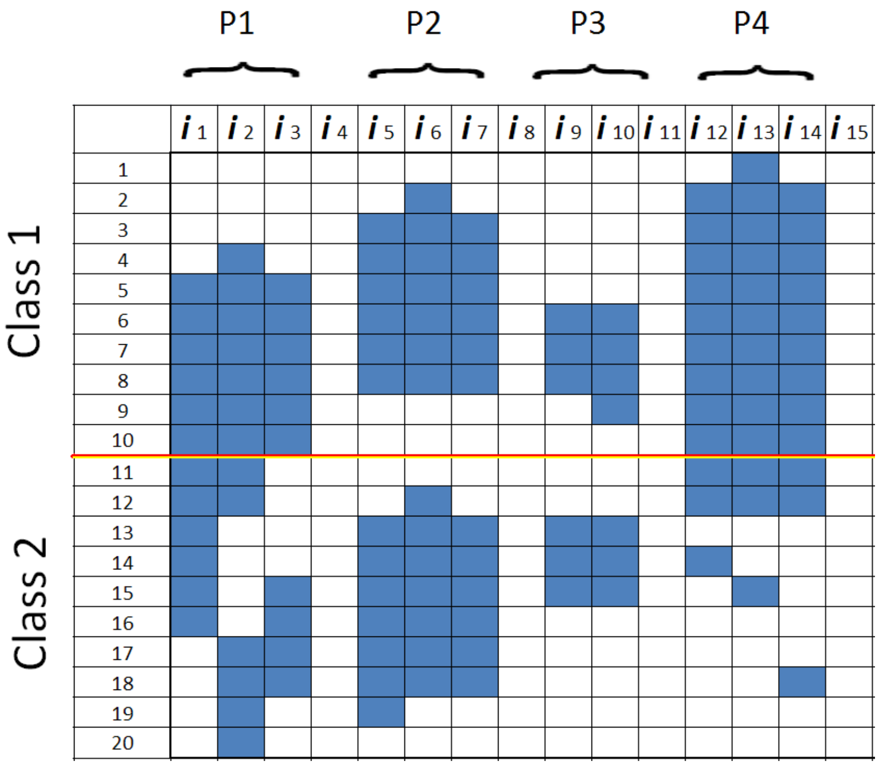


Fig. 1. A sample data set with interesting discriminative patterns (P_1, P_4) and uninteresting patterns (P_2, P_3)

Note that the discriminative measures discussed above are generally not anti-monotonic as shown by [14], [5], [9]. Take *DiffSup* for instance (while other measures like support ratio, information gain and odds ratio are not anti-monotonic either): although the *DiffSup* of the three items in P_1 are 0, 0 and 0.2 respectively, P_1 has a *DiffSup* of 0.6 as an itemset. Due to the lack of anti-monotonicity, these measures can not be directly used in an Apriori framework [2] for exhaustive and efficient pattern mining as can be done for measures like support[2], h-confidence [53] etc. To address this issue, many approaches [29], [28], [55], [11], [9] adopt a two-step strategy (denoted as Group A), where first, a frequent pattern mining algorithm is used to find all (closed) frequent patterns that satisfy a certain support threshold $minsup$ either from the whole dataset or from only one of the classes. The patterns found can be further refined using other interestingness measures (e.g. [7], [23], [44]). Then, as post-processing, *DiffSup* is computed for each of these patterns, based on which discriminative patterns are selected. Note that, in general, these two-step approaches can work even with a very low $minsup$ threshold [49], [9] on relatively sparse or low-dimensional datasets.

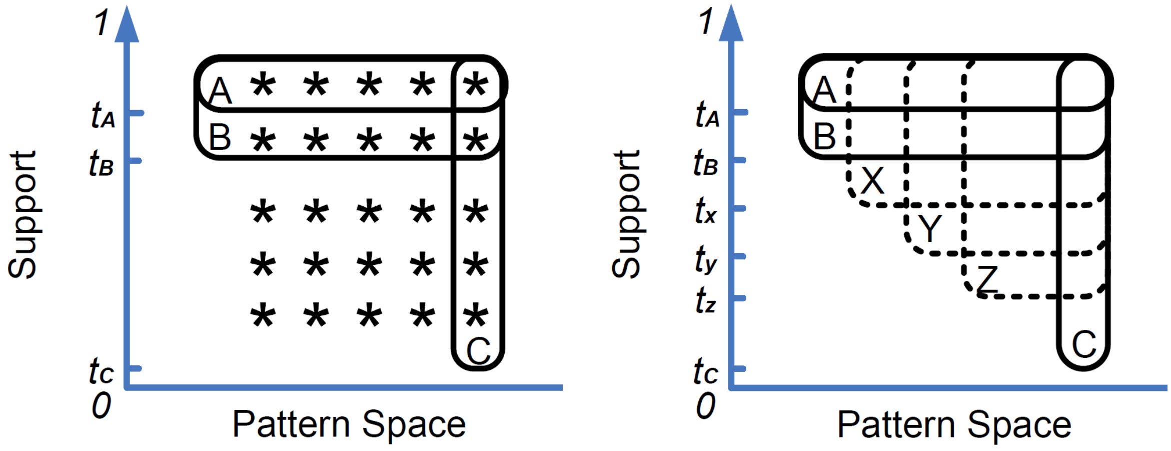
However, since these approaches ignore class label information in the mining process, many frequent patterns discovered in the first step may turn out to have low discriminative power in the second step. For instance, in Figure 1, the relative supports of P_2 and P_3 in the whole dataset are 0.6 and 0.3 respectively, and will be considered as frequent patterns if the support threshold is 0.2. However, P_2 and P_3 are not discriminative since they both have a $DiffSup$ of 0. In particular, in datasets with relatively high density [4] and high-dimensionality, a huge number of non-discriminative patterns like P_2 and P_3 in Figure 1 may exist. Such patterns may meet the $minsup$ threshold and would be discovered in the first step, but would be found as non-discriminative patterns in the second step. If a low $minsup$ is used, a huge number of such patterns can reduce the efficiency of both the two steps as discussed in [10]. In such a situation, the two-step approaches have to use a sufficiently high $minsup$ in order to generate the complete set of results within an acceptable amount of time, and thus may miss a large number of highly discriminative patterns that fall below the $minsup$ threshold.

A possible strategy for improving the performance of the two-step approaches is to directly utilize the support of a pattern in the two classes for pruning some non-discriminative patterns in the pattern mining stage. Indeed, several approaches have been proposed [5], [9], where the anti-monotonic upper bounds of discriminative measures, such as $DiffSup$, are used for pruning some non-discriminative patterns in an Apriori-like framework [2]. This strategy, like the two-step approaches, also guarantees to find the complete set of discriminative patterns with respect to a threshold, although in a more efficient manner. However, in datasets with relatively high density and high-dimensionality, there can be a large number of frequent non-discriminative patterns like P_2 in Figure 1. Such patterns may not be pruned by these approaches because the upper bounds of the discriminative measures may be weak (technical details in Section III). Thus, as illustrated in Figure 2(a), these approaches (referred to as group B in the rest of this paper) are able to discover a larger fraction of the interesting discriminative patterns as compared to the two-step approaches. However, they may still miss a lot of highly discriminative patterns, particularly those at low support levels, given the same fixed amount of time. These low support patterns are supported by a relatively small number of samples but can still be highly discriminative according to their $DiffSup$ value, especially in the case of data sets with skewed class size distributions.

[4] The density of a transaction matrix is the percentage of $1's$ in the transaction-by-item matrix

Yet another strategy for discovering a significant subset of the discriminative patterns is to directly use a measure of discriminative power for pruning non-discriminative patterns [56]. As an instance of such an approach, $\mathit{DiffSup}$ can be computed for each candidate pattern α, and if $\mathit{DiffSup}(\alpha) < r$, then α and all its supersets can be pruned in an Apriori-like algorithm [2]. This strategy is computationally more efficient than the two-step approaches, because no patterns with $\mathit{DiffSup}(\alpha) < r$ are generated during the mining process. However, this improved efficiency comes at the cost of not discovering the complete set of discriminative patterns, since $\mathit{DiffSup}$ is not anti-monotonic [5]. More specifically, the algorithms in this group (referred to as group C in the rest of this paper), may miss interesting discriminative patterns whose individual items are not discriminative (e.g. P_1 in Fig. 1). With respect to the coverage of the set of interesting discriminative patterns, the approaches in this group may be able to discover low-support patterns at the expense of missing a large number of interesting patterns, as illustrated by the stars not included in box C in Figure 2(a). This observation is also reflected in our experimental results (section VI-B.2).

As can be seen from the discussion above, which is summarized in Figure 2(a), the current approaches face an inherent trade-off when discovering discriminative patterns from a dense and high-dimensional data set. The approaches in groups A and B face challenges with discovering low-support patterns due to their focus on the complete discovery of discriminative patterns satisfying the corresponding thresholds. On the other hand, the approaches in group C sacrifice completeness for the ability of discovering low-support discriminative patterns. This trade-off is expected to be faced by any algorithm for this complex problem, particularly due to the restriction of fixed computational time. In such a scenario, an appropriate approach to discover some of the interesting discriminative patterns missed by the current approaches, is to formulate new measures for discriminative power and corresponding algorithms that can progressively explore lower support thresholds for discovering patterns, while trading off completeness to some extent. Such a design is illustrated in Figure 2(b), where boxes X, Y and Z represent three approaches, which can discover patterns with progressively lower thresholds ($t_x > t_y > t_z$). However, the cost associated with this ability is that of potentially missing some patterns that are at higher support levels. Still, X, Y and Z can all discover several patterns that are exclusive to only one of them, and can thus play a complementary role to the existing approaches by expanding the coverage of the set of interesting discriminative patterns.



(a) The limitation of existing approaches

(b) The motivation of the proposed work

Fig. 2. An illustration of the coverage of the space of discriminative patterns by different approaches given the same amount of time. The t's on the y-axis represent the lowest support of the patterns that are respectively covered by the corresponding approaches (represented by boxes), given the same and fixed amount of time. (a) Box A, B and C represent the set of patterns discovered by the corresponding approaches in group A, B and C respectively. (b) Illustration of the tradeoff between the capability to search low-support discriminative patterns in dense and high-dimensional data and the completeness of the pattern discovery. Boxes X, Y and Z represent three conceptual low-support discriminative pattern mining approaches that discover patterns not found by the approaches in groups A, B and C. Note that, in this figure, the set of interesting discriminative patterns is the same as that in (a), but the corresponding $*$'s are not shown for the sake of clarity.

Corresponding to the motivation discussed above, we propose a family of anti-monotonic measures of discriminative power named $SupMaxK$. These measures conceptually organize the set of discriminative patterns into nested layers of subsets, which are progressively complete in their coverage, but require increasingly more computation for their discovery. Essentially, $SupMaxK$ estimates the $DiffSup$ of an itemset by calculating the difference of its support in one class and the maximal support among all of its size-K subsets in the other class. The smaller the value of K, the more effective $SupMaxK$ is for finding low-support discriminative patterns by effectively pruning frequent non-discriminative patterns. Notably, due to the anti-monotonicity property of all the members of $SupMaxK$, each of them can be used in an Apriori-like framework [2] to guarantee the discovery of all the discriminative patterns with $SupMaxK$ $\geq r$, where r is a user-specified threshold. Given the same (limited) amount of time, the members of this family provide a tradeoff between the ability to search for low-support discriminative patterns and the coverage of the space of valid discriminative patterns for the corresponding threshold, as illustrated by the three conceptual approaches X, Y and Z in Figure 2(b). In particular, we find that a special member with $K = 2$ named $SupMaxPair$, is suitable for dense and high-dimensional data. We have designed a framework, named SMP, which uses $SupMaxPair$ for discovering discriminative patterns. Carefully designed experiments with both synthetic datasets and a cancer gene expression dataset are used to demonstrate that SMP can serve a complementary role to the existing approaches by discovering low-support yet highly discriminative patterns from dense and high-dimensional data, when the latter fail to discover them within an acceptable amount of time.

A. Contributions of this paper

The contributions of this paper can be summarized as follows:

1) We address the necessity of trading off the completeness of discriminative pattern discovery with the ability to discover low-support discriminative patterns from dense and high-dimensional data within an acceptable amount of time. For this, We propose a family of anti-monotonic measures named $SupMaxK$ that conceptually organize the set of discriminative patterns into nested layers of subsets, which are progressively more complete in their coverage, but require increasingly more computation for their discovery.
2) In particular, $SupMaxK$ with $K = 2$, named $SupMaxPair$, is a special member of this family that is suitable for dense and high-dimensional data, and can serve a complementary role to the existing approaches by helping to discover low-support discriminative patterns, when the latter fail to discover them within an acceptable amount of time. We designed a framework, named SMP, which uses $SupMaxPair$ for discovering discriminative patterns.
3) A variety of experiments with both synthetic datasets and a cancer gene expression dataset are presented to demonstrate that there are many patterns with relatively low support that can be discovered by SMP but not by the existing approaches. In particular, these experiments rigorously demonstrate that the low-support discriminative patterns discovered only by SMP from the cancer gene expression dataset are statistically significant (via permutation test [18], [42]) and biologically relevant (via comparison with a list of cancer-related genes [21] and a collection of biological gene sets [42] (e.g. pathways)). These are the recognized methods for evaluating the utility of such patterns for applications such as biomarker discovery [42], [8], [22].

The source codes and dataset used in this paper are available at http://vk.cs.umn.edu/SMP/.

II. Basic Terminology and Problem Definition

Let D be a dataset with a set of m items, $I = \{i_1, i_2, ..., i_m\}$, two class labels S_1 and S_2, and a set of n labeled instances (itemsets), $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, where $\boldsymbol{x}_i \subseteq I$ is a set of items and $y_i \in \{S_1, S_2\}$ is the class label for $\boldsymbol{x}_i$. The two sets of instances that respectively belong to the class S_1 and S_2 are denoted by D^1 and D^2, and we have $|D| = |D^1| + |D^2|$. For an itemset $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_l\}$ where $\alpha \subseteq I$, the set of instances in D^1 and D^2 that contain α are respectively denoted by D^1_α and D^2_α. The relative supports of α in classes S_1 and S_2 are $RelSup^1(\alpha) = \frac{|D^1_\alpha|}{|D^1|}$ and $RelSup^2(\alpha) = \frac{|D^2_\alpha|}{|D^2|}$, respectively. $RelSup$ is anti-monotonic since the denominator is fixed and the numerator is support of the itemset, which is anti-monotonic.

The absolute difference of the relative supports of α in D^1 and D^2 is defined originally in [5] and denoted in this paper as $DiffSup$:

$$DiffSup(\alpha) = |RelSup^1(\alpha) - RelSup^2(\alpha)|. \tag{1}$$

An itemset α is $r - discriminative$ if $DiffSup(\alpha) \geq r$. The problem addressed by discriminative pattern mining algorithms is to discover all patterns in a dataset with $DiffSup \geq r$.

Without loss of generality, we only consider discriminative patterns for the binary-class problem. Our work can be extended to multiple classes as described in [5].

III. Computational Limitations of Current Approaches

As discussed in Section I, in dense and high-dimensional data, the approaches in groups A and B have to use a relatively high threshold in order to provide the complete result within an acceptable amount of time. In this section, we will show that this limitation is essentially due to the ineffective pruning of frequent non-discriminative patterns (e.g. P_2 in Figure 1). Generally, the approaches in group B is relatively more efficient than those in group A, as discussed in Section I. We use the measure originally proposed in CSET [5] as a representative of group B for this discussion, while a similar discussion also holds for other approaches in group B [9], [34]. In CSET, an upper bound of $DiffSup$ is defined as the bigger of the relative supports of a pattern α in D^1 and D^2. In this paper, we denote it as $BiggerSup$:

$$BiggerSup(\alpha) = max(\mathit{RelSup}^1(\alpha), \mathit{RelSup}^2(\alpha)). \tag{2}$$

Lemma 1. BiggerSup is anti-monotonic

Proof. Follows from the anti-monotonicity of RelSup and the property of the max function. □

Since $\mathit{BiggerSup}$ is an upper bound of $\mathit{DiffSup}$ [5], and it is also anti-monotonic (Lemma 1), CSET [5] uses $\mathit{BiggerSup}$ as a pruning measure in a Apriori-like framework, and can discover, given sufficient time and computing resources, the complete set of discriminative patterns (w.r.t a $\mathit{BiggerSup}$ threshold). However, by using the bigger one to estimate the difference of the two supports, $\mathit{BiggerSup}$ is a weak upper bound of $\mathit{DiffSup}$. For instance, if we want to use CSET to search for $0.4 - discriminative$ patterns in Figure 1, P_3 can be pruned, because it has a $\mathit{BiggerSup}$ of 0.3. However, P_2 can not be pruned ($\mathit{BiggerSup}(P_2) = 0.6$), even though it is not discriminative ($\mathit{DiffSup}(P_2) = 0$). More generally, $\mathit{BiggerSup}$-based pruning can only prune infrequent non-discriminative patterns with relatively low support, but not frequent non-discriminative patterns. Therefore, in dense and high-dimensional data, where a large number of frequent non-discriminative patterns are expected to exist, CSET with a relatively low $\mathit{BiggerSup}$ threshold can often fail to produce the complete results in a reasonable amount of time. Thus, CSET has to set the $\mathit{BiggerSup}$ threshold high and may not discover discriminative patterns at lower support that may be of interest. Similar discussion on the limited ability of pruning frequent non-discriminative patterns also holds for other approaches in groups A and B, i.e., all the two-step approaches, and those based on the information gain upper bound [9], and other statistical metric-based pruning [5], [34].

IV. Proposed Approach

As shown above, the limitation of existing approaches is essentially the ineffectiveness of pruning frequent non-discriminative patterns. Conceptually, to prune frequent non-discriminative patterns, a new measure should be designed such that a pattern's support in one class can be effectively limited to a relatively smaller number compared to its support in the other class. In this section, we start with such a measure $\mathit{SupMax1}$ in Definition 1, and then extend it to a family of measures $SupMaxK$. Then, we will discuss the relationships between $\mathit{DiffSup}$, $\mathit{BiggerSup}$ and $SupMaxK$. Finally, we will focus on a special member of this family $\mathit{SupMaxPair}$ that is suitable for high-dimensional data. Note that, for an itemset α two cases can happen: $\mathit{RelSup}^1(\alpha) \geq \mathit{RelSup}^2(\alpha)$ or $\mathit{RelSup}^1(\alpha) < \mathit{RelSup}^2(\alpha)$. In the following discussion, without loss of generality, we assume $\mathit{RelSup}^1(\alpha) \geq \mathit{RelSup}^2(\alpha)$ for simplicity.

A. SupMax1: A Simple Measure to Start with

Definition 1. *The* $\mathit{SupMax1}$ *of an itemset* α *in* D^1 *and* D^2 *is defined as*

$$SupMax1(\alpha) = \mathit{RelSup}^1(\alpha) - max_{a\in\alpha}(\mathit{RelSup}^2(\{a\})).$$

$\mathit{SupMax1}$ of an itemset α is computed as the difference between the support of α in D^1, and the maximal individual support of the items in α in D^2. $\mathit{SupMax1}$ approximates $\mathit{DiffSup}$ by using the maximal individual support in D^2 to estimate $\mathit{RelSup}^2(\alpha)$. Clearly, the maximal individual support is quite a rough estimator for $\mathit{RelSup}^2(\alpha)$, because a pattern can have very low support in class S_2 but the items in it can still have very high individual supports in this class. However, an alternative way to interpret $\mathit{SupMax1}$ is that, a pattern with large $\mathit{SupMax1}$ has relatively high support in one class and all the items in it have relatively low support in the other class. P_4 is such an example whose $\mathit{SupMax1}$ is $0.9 - max(0.3, 0.3, 0.3) = 0.6$ as shown in Figure 1. Thus, given a $\mathit{SupMax1}$ threshold, say 0.4, $\mathit{SupMax1}$ discovers a subset of $0.4 - discriminative$ patterns but not all, e.g. it will miss patterns like P_1 in Figure 1, which has relatively high $\mathit{DiffSup}$ (0.6) but zero $\mathit{SupMax1}$.

B. SupMaxK

Following the rationale of $\mathit{SupMax1}$, the maximal support of size-k subsets of a pattern in D^2 can be used to estimate $\mathit{RelSup}^2(\alpha)$ instead of using maximal individual support in class S_2 to estimate $\mathit{RelSup}^2(\alpha)$. This can provide a better estimation of $\mathit{RelSup}^2(\alpha)$. In such a manner, $\mathit{SupMax1}$ can be generalized into a family of measures $SupMaxK$, which is formally defined in Definition 2. Note that in the following discussion, $SupMaxK$ will be used to refer to this family as well as one of its general members, for the clarity of presentation.

Definition 2. *The* $SupMaxK$ *of an itemset* α *in* D^1 *and* D^2 *is defined as*

$$SupMaxK(\alpha) = \mathit{RelSup}^1(\alpha) - max_{\beta\subseteq\alpha}(\mathit{RelSup}^2(\beta)),\ \textit{where } |\beta| = K$$

So, $SupMaxK$ of an itemset α is computed as the difference between the support of α in D^1, and the maximal support among all the size-K subsets of α in D^2. Note that, in this paper, $SupMaxK$ is defined with respect to $DiffSup$, while similar concept can also be applied to other discriminative measures such as the ratio-based measure [14].

C. Properties of the $SupMaxK$ Family

In the following subsections, we discuss three properties of the $SupMaxK$ family.

1) The subset-superset relationship among SupMaxK members: Based on the definition of $SupMaxK$, the following two Lemmas show the relationship among $SupMaxK$ members.

Lemma 2. *If we use $MaxSup(\alpha, K)$ to denote the second component of $SupMaxK(\alpha)$, i.e. $max_{\beta \subseteq \alpha}(RelSup(\beta))$ with $|\beta| = K$, then $MaxSup(\alpha, K)$ is a lower bound of $MaxSup(\alpha, K - 1)$ for integer $K \in [2, |\alpha|]$*

Proof. For every size-$(K - 1)$ subset of α (say β, $|\beta| = K - 1$), there exists a size-K subset of α (say β', $|\beta'| = K$) such that $\beta \subset \beta'$, e.g. by adding any i to β, where $i \in \alpha$ and $i \notin \beta$. Based on the anti-monotonicity property of $RelSup$, it is guaranteed that $RelSup(\beta') \leq RelSup(\beta)$. Then, from the properties of the max function, $max_{\beta' \subseteq \alpha}(RelSup(\beta')) \leq max_{\beta \subseteq \alpha}(RelSup(\beta))$. Thus, $MaxSup(\alpha, K)$ is a lower bound of $MaxSup(\alpha, K - 1)$. □

Lemma 3. *$SupMax(K - 1)$ of an itemset α is a lower bound of its $SupMaxK$, or alternatively $SupMaxK$ of an itemset α is an upper bound of its $SupMax(K - 1)$, for integer $K \in [2, |\alpha|]$*

Proof. Follows directly from Definition 2, Lemma 2. □

From Lemma 3, we know that, given the same threshold r and sufficient time, the set of patterns discovered with $SupMax(K - 1)$ in an Apriori framework is a subset of the set of patterns discovered with $SupMaxK$. This means that $SupMaxK$ can find more and more discriminative patterns as K increases from 1 ($SupMax1$), to 2 (*SupMax2*), to 3 (*SupMax3*) and so on. The patterns that are discovered by $SupMaxK$ but not by $SupMax(K - 1)$ are those with $SupMaxK \geq r$, but with $SupMax(K - 1) < r$. Figure 3 shows an extended version of the data set shown in Figure 1 containing fifteen addition items ($i_{16} - i_{30}$) and two patterns P_5 and P_6, the rest being identical to Figure 1. In this data set, given the same threshold $r = 0.4$, $SupMax1$

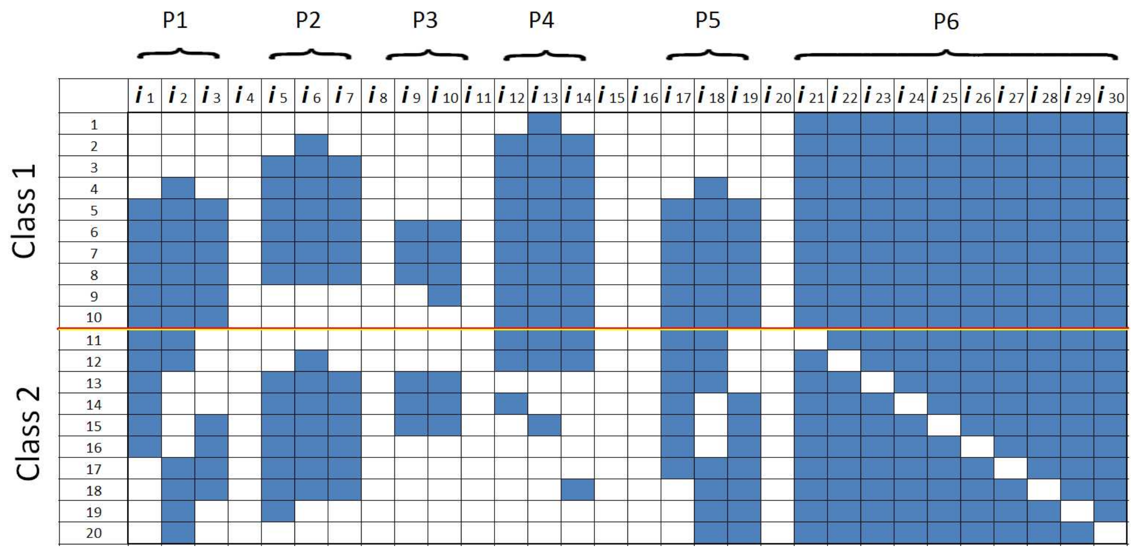

P1
P2
P3
P4
P5
P6
i 1 i 2 i 3 i 4 i 5 i 6 i 7 i 8 i 9 i 10 i 11 i 12 i 13 i 14 i 15 i 16 i 17 i 18 i 19 i 20 i 21 i 22 i 23 i 24 i 25 i 26 i 27 i 28 i 29 i 30
Class 1
Class 2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Fig. 3. An extended version of the data set shown in Figure 1 containing fifteen addition items ($i_{16} - i_{30}$) and two patterns P_5 and P_6, the rest being identical to Figure 1.

can find P_4, but not P_1 and P_5, both of which have $\mathit{DiffSup} = 0.6$, but zero $\mathit{SupMax1}$; $\mathit{SupMax2}$ can find P_1 in addition to P_4; furthermore, $\mathit{SupMax3}$ can find P_5 in addition to P_4 and P_1. This illustrates that $\mathit{SupMax3}$ can find all the patterns found using $\mathit{SupMax1}$ and $\mathit{SupMax2}$, but not vice versa, as discussed above. Furthermore, ***SupMax10*** will be able to discover pattern P_6 in addition to the patterns found using $\mathit{SupMax1}$, $\mathit{SupMax2}$ and $\mathit{SupMax3}$.

2) The Exactness of the $\mathit{SupMaxK}$ Family: Lemmas 2 and 3 lead to Theorem 1, which shows the relationship between $\mathit{SupMaxK}$ and $\mathit{DiffSup}$.

Theorem 1. *$\mathit{SupMaxK}$ is a lower bound of $\mathit{DiffSup}$, for integer $K \in [1, |\alpha| - 1]$.*

Proof. Since ***DiffSup(α)*** is equivalent to ***SupMaxK(α)*** with $K = |\alpha|$ (We assumed $\mathit{RelSup}^1(\alpha) \geq \mathit{RelSup}^2(\alpha)$ for simplicity earlier this Section), this theorem follows from Lemma 3. □

Theorem 1 guarantees that the patterns discovered by any $\mathit{SupMaxK}$ members with threshold r also have $\mathit{DiffSup} \geq r$. Therefore, $\mathit{SupMaxK}$ members with threshold r discover only $r - \mathit{discriminative}$ patterns.

3) The Increasing Completeness of the $\mathit{SupMaxK}$ Family: The max function together with the anti-monotonicity of RelSup yields the following result about the anti-monotonicity of each member of $\mathit{SupMaxK}$.

Theorem 2. *Each member of $SupMaxK$ is anti-monotonic.*

Proof. Let $\alpha \subseteq I$ be an itemset, and $\alpha' \subseteq I$ be a superset of α, such that $\alpha' = \alpha \cup \{i\}$, where $i \in I$ and $i \notin \alpha$. Firstly, from the anti-monotonicity of $RelSup$, we have $RelSup^1(\alpha') \leq RelSup^1(\alpha)$. Then, based on the property of the max function, $max_{\beta' \subseteq \alpha'}(RelSup^2(\beta')) \geq max_{\beta \subseteq \alpha}(RelSup^2(\beta))$, where $|\beta| = K$ and $|\beta'| = K$. Finally, we have the following:

$$\begin{aligned} SupMaxK(\alpha') &= RelSup^1(\alpha') - max_{\beta' \subseteq \alpha'}(RelSup^2(\beta')) \\ &\leq RelSup^1(\alpha) - max_{\beta \subseteq \alpha}(RelSup^2(\beta)) \\ &= SupMaxK(\alpha). \end{aligned}$$

□

Based on Theorem 2, given a threshold r, any member of the $SupMaxK$ family can be used within an Apriori-like framework [2] to discover the complete set of patterns with $SupMaxK \geq r$. Note that $SupMaxK$ could be alternatively defined using the min function, thus providing a better estimation of $DiffSup$. However, this version of $SupMaxK$ will not be anti-monotonic and thus can not be used in the Apriori framework for the systematic search of discriminative patterns.

Since there are a finite number of discriminative patterns in a dataset given a $DiffSup$ threshold, and $SupMaxK$ finds more and more discriminative patterns as K increases (Lemma 3), the set of patterns discovered with $SupMaxK$ and threshold r within an Apriori-like framework is increasingly more complete with respect to the complete set of $r-discriminative$ patterns.

4) Summary of the three properties of the SupMaxK Family: From the subset-superset relationship among $SupMaxK$ members, and the exactness and increasing completeness of the $SupMaxK$ family, $SupMaxK$ members conceptually organize the complete set of discriminative patterns into nested subsets of patterns that are increasingly more complete in their coverage with respect to $r-discriminative$ patterns. This yields interesting relationships between $DiffSup$, $BiggerSup$ and the $SupMaxK$ family, which are discussed below.

D. Relationship between DiffSup, BiggerSup and the SupMaxK Family

To understand relationship among $DiffSup$, $BiggerSup$ and $SupMaxK$, Figure 4 displays the nested structure of the $SupMaxK$ family together with $DiffSup$ and $BiggerSup$ from the perspective of the search space of discriminative patterns in a dataset. L_{All} is the complete set of $r-discriminative$ patterns given a $DiffSup$ threshold r. L_{CSET} is the search space explored by CSET in order to find all the patterns in L_{All}. Note that L_{CSET} is a superset of L_{All}, because $BiggerSup$ is an upper bound of $DiffSup$. Note that, L_{CSET} can be much larger than L_{All} for dense and high-dimensional data sets, especially when a relatively low $BiggerSup$ threshold is used. In such cases, CSET may not be able to generate complete results within an acceptable amount of time. For instance, on the cancer gene expression data set used in our experiments, the lowest $BiggerSup$ threshold for which CSET can produce the complete results within 4 hours is 0.6. With a lower threshold 0.4, CSET can not produce the complete results within 24 hours.

Members of the $SupMaxK$ family help address this problem with $BiggerSup$ by stratifying all the $r-discriminative$ patterns into subsets that are increasingly more complete (Set L_1, L_2, $\dots$, L_k, L_{k+1}, $\dots$, L_{All}), as shown in Lemma 3 and the subsequent discussion, and illustrated in Figure 4. However, note that these superset-subset relationships among $SupMaxK$ members and between $SupMaxK$ and $BiggerSup$ (used by CSET) hold only when the same threshold is used for $BiggerSup$ and all the $SupMaxK$ members, and unlimited computation time is available. In practice, given the same fixed amount of time, progressively lower thresholds can be used for $SupMaxK$ members as K decreases. This tradeoff was illustrated earlier in Figure 2(b).

Since the focus of this paper is on dense and high-dimensional data, another practical factor should be considered, that is, the computational efficiency of the $SupMaxK$ members. In the next section, we will introduce a special member of the $SupMaxK$ family that is computationally suitable for dense and high-dimensional data.

E. SupMaxPair: A Special Member Suitable for High-Dimensional Data

In the previous discussion, we showed that as K increases, the set of patterns discovered with $SupMaxK$ and threshold r in an Apriori framework is increasingly more complete with respect to the complete set of $r - discriminative$ patterns. Thus, in order to discover as many $r - discriminative$ patterns as possible, an as large as possible value of K should be used given the time limit. However, it is worth noting that the time and space complexity to compute and store the second component in the definition of $SupMaxK$, i.e. $MaxSup(\alpha, K) = max_{\beta \in \alpha}(RelSup^2(\beta))$ with $|\beta| = K$ are both $O(m^K)$ (The exact times of calculation is $\binom{M}{K}$), where M is the number of items in the dataset. In high-dimensional data set (large M), $K > 2$ is usually infeasible. For instance, if there are 10000 items in the data set ($M = 10000$), even $SupMaxK$ with $K = 3$ will require the computation of the support of all $\binom{10000}{3} \approx 1.6 \times 10^{11}$

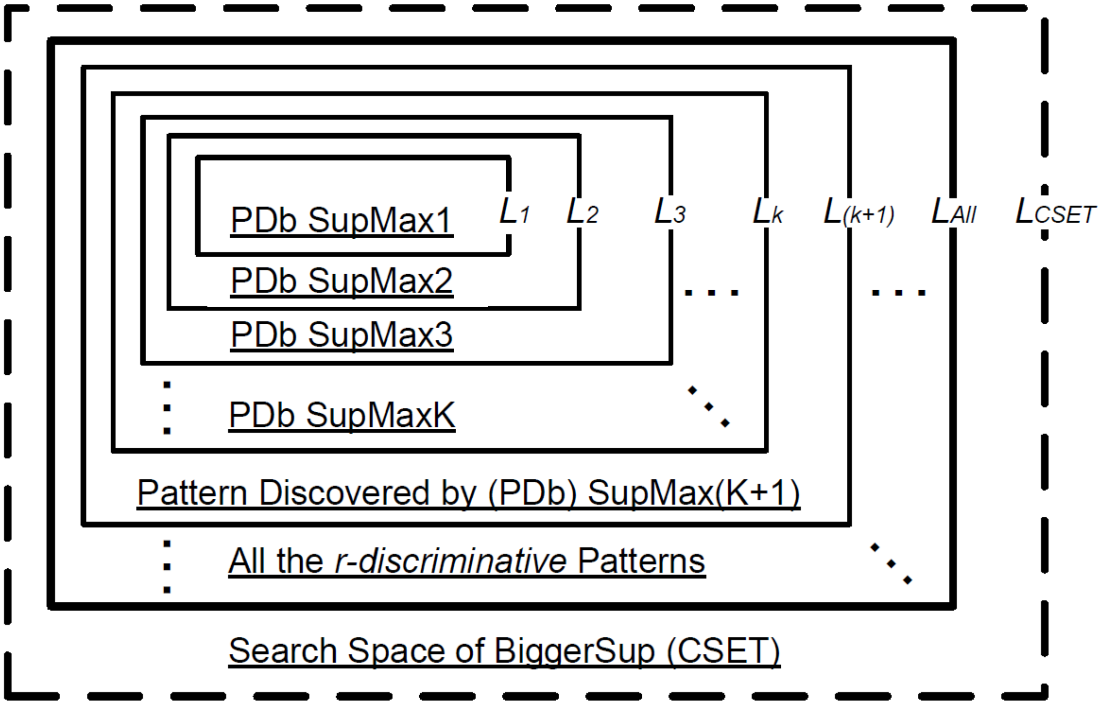

PDb SupMax1
PDb SupMax2
PDb SupMax3
PDb SupMaxK
Pattern Discovered by (PDb) SupMax(K+1)
All the r-discriminative Patterns
Search Space of BiggerSup (CSET)
L1
L2
L3
Lk
L(k+1)
LAll
LCSET

Fig. 4. Nested layers $(L_1, L_2, L_3, \ldots, L_k, L_{k+1}, \ldots, L_{All}, L_{CSET})$ of patterns defined by $\mathit{SupMaxK}$, and relationship with the complete set of discriminative patterns (layer L_{All}), and the search space of $\mathit{BiggerSup}$ used by CSET (layer L_{CSET}). (PDb stands for ”Patterns Discovered by”.) $PDbSupMaxK$ is a subset of $PDbSupMax(K+1)$. Note that this figure only shows the subset-superset relationship, while the size of each rectangle does not the imply number of patterns in each set.

size-3 patterns. Therefore, due to our emphasis on dense and high-dimensional data, we will focus on $\mathit{SupMaxK}$ with $K = 2$, i.e., $\mathit{SupMaxPair}$, to balance the accurate estimation of $\mathit{DiffSup}$ and computational efficiency. Note that, based on the definition of $\mathit{SupMaxPair}$, the computational complexity of the second component of $\mathit{SupMaxPair}$ (maximal pair-wise support in class S_2) for an itemset $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_l\}$ with size greater than 2 is $O(l^2)$. However, according to the Apriori framework [2], $\mathit{MaxSup}(\alpha, \mathit{2})$ only depends on three terms that will have been computed before the computation of $\mathit{MaxSup}(\alpha, \mathit{2})$ itself: $\mathit{MaxSup}(\{\alpha_1, \alpha_2, ..., \alpha_{l-1}\}, 2)$ and $\mathit{MaxSup}(\{\alpha_1, \alpha_2, ..., \alpha_{l-2}, \alpha_l\}, 2)$, and $\mathit{MaxSup}(\{\alpha_{l-1}, \alpha_l\}, 2)$, and thus the computational complexity for $\mathit{MaxSup}(\alpha, \mathit{2})$ is $O(1)$ per itemset α.

As shown in Figure 4, $\mathit{SupMaxPair}$ can perform a complete search of the $r-discriminative$ patterns in the first two layers, even for a low value of r. Indeed, we will demonstrate in our experimental results on a cancer gene expression data set (Section VI) that searching these two layers itself can enable SupMaxPair to discover many low-support patterns that may not be discovered by CSET within an acceptable amount of time. Furthermore, these patterns are statistically significant and biologically relevant.

Before we discuss these results, we lay out the complete framework that we use for discovering discriminative patterns from dense and high-dimensional data.

V. Framework for Discriminative Pattern Mining

In this section, we explain the major steps in the framework used for discriminative pattern mining in our experiments:

• **Step 1**: This is an algorithm-specific step. For example, for $SupMaxPair$, all the item-pair supports are computed and stored in a matrix, whose (i, j) entry is the item-pair support of item i and j. The complexity of this step is $O(nm^2)$, where n is the number of transactions, and m is the number of unique items. No such pre-computation has to be done for CSET.

• **Step 2**: The Apriori framework [2] is used in this step for discriminative pattern mining using the anti-monotonic measures $BiggerSup$ and $SupMaxPair$. For SMP, discriminative patterns are firstly mined from one class and then mined from the other, while CSET discovers patterns once from the whole dataset.

• **Step 3**: To facilitate further pattern processing and pattern evaluation, we selected only the closed itemsets [37] from the complete set of itemsets produced.

For clarity, we refer to the version of this framework where $BiggerSup$ is used for discovering patterns as **CSET**, while the version using $SupMaxPair$ is referred to as **SMP** in the subsequent discussion. Our analysis of the quality of the patterns and the computational time requirements are presented with respect to the patterns produced by these complete pipelines.

VI. Experimental Results

In order to evaluate the efficacy of different discriminative pattern mining algorithms, particularly CSET (a representative of the approaches in group B discussed in Section I) and our proposed algorithm SMP, we designed two sets of experiments. The first set of experiments utilize synthetic data sets with varying density and dimensionality to study the properties of CSET and SMP. The second set of experiments involve the application of CSET and SMP to a breast cancer gene expression data. The second set aims at a systematic evaluation of the statistical significance and biological relevance of the resultant patterns, thus validating the effectiveness of CSET and SMP for knowledge discovery from real data. All the experiments presented here were run on a Linux machine with 8 Intel(R) Xeon(R) CPUs (E5310 @ 1.60GHz) and 16GB memory.

A. Experiments on Synthetic Data Sets with Varying Density and Dimensionality

In the first set of experiments, we study the performance of SMP and CSET on synthetic binary data sets whose background can be fully controlled. Specifically, we created two collections of synthetic datasets respectively with (i) varying density and fixed dimensionality, and (ii) varying dimensionality and fixed density. We first describe the approach we used to create these two collections of data sets and then present the performance of SMP compared to CSET.

1) Methodology for Generating Synthetic Data Sets: Each synthetic data set have two major components: discriminative and non-discriminative patterns. Discriminative patterns are the target of the mining algorithms, while non-discriminative patterns are obstacles. As discussed in Section I, an effective discriminative pattern mining algorithm should be able to prune the non-discriminative patterns at early stage while discovering discriminative patterns.

Ten discriminative patterns each of sizes 2, 4, 6, 8 and 10 were embedded in each synthetic dataset, resulting in a total of 50 discriminative patterns per dataset. To reflect the distribution of different types of discriminative patterns in real data, for each of the five sizes, we randomly determined a number of patterns (out of ten) that can be discovered by CSET but not SMP (type-I), and the remaining patterns that can be discovered by SMP but not CSET (type-II). Specifically, type-I patterns are those that have $DiffSup$ greater than 0.2, but $SupMaxPair$ below 0.2. As discussed in Section IV, SMP can not find type-I patterns due to the fact that $SupMaxPair$ is an lower bound of $DiffSup$. In contrast, type-II patterns are those that have $BiggerSup$ below the lowest threshold (0.2) that CSET can finish within an acceptable amount of time (we use 4 hours as the representative acceptable amount of time). SMP can find these type-II patterns if it can effectively prune non-discriminative patterns and can search at lower support levels (0.1). Table I displays the number of type-I and type-II discriminative patterns of different sizes embedded in each of the synthetic datasets. Note that these numbers are kept the same for all the synthetic datasets to ensure that results across across different datasets are comparable. Note that in practice, there may be other types of patterns that can be discovered by both CSET and

	size-2	size-4	size-6	size-8	size-10
type-I patterns	3	6	5	8	7
type-II patterns	7	4	5	2	3
Total patterns of each size	10	10	10	10	10

TABLE I

NUMBER OF TYPE-I AND TYPE-II DISCRIMINATIVE PATTERNS OF SIZE-2, 4, 6, 8 AND 10.

SMP. In this analysis, we do not embed these other types of patterns and focus only on the effectiveness of CSET and SMP for discovering different types of discriminative patterns.

For all the synthetic data sets, we fix the number of samples at 700, in which half are of class 1 and the other half are of class 2. Two collections of datasets were generated as follows.

Varying density with fixed dimensionality: For this collection of data sets, we fix the dimensionality at 4000. After we embed the 50 discriminative patterns, we have the first dataset of density 10%. Next, we keep adding non-discriminative patterns of size 10 and support greater than 0.2, and create four more data sets with densities of 0.13, 0.16, 0.19 and 0.22 respectively.

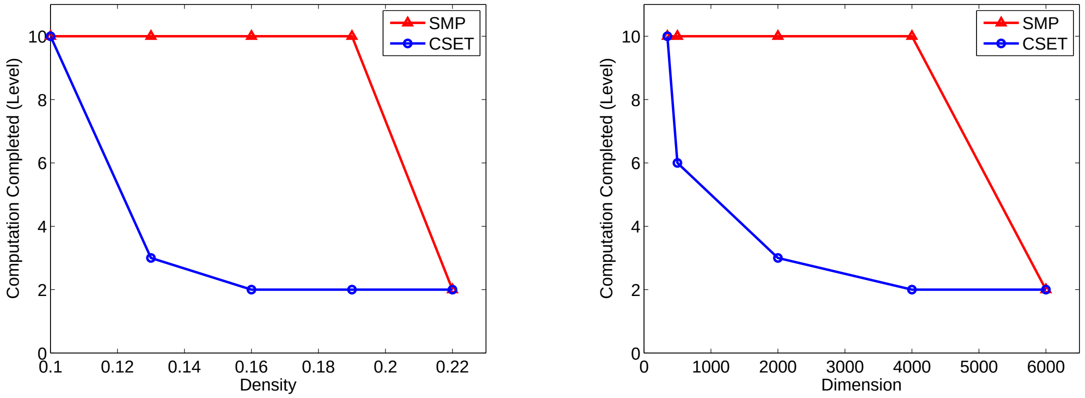
Varying dimensionality with fixed density: For this collection of data sets, we fix the density of the dataset at 0.2. After we embed the 50 discriminative patterns (density 10%), we further add non-discriminative patterns to make the density equal to 0.2 and use this dataset as the first dataset (the dimensionality is 350). Next, we further add non-discriminative patterns of size-10 and support greater than 0.2 and simultaneously increase the dimensionality of the data set to maintain the density at 0.2. In this way, we create another four data sets with dimensionalities of 500, 2000, 4000 and 6000.

Note that the supporting transactions of both the discriminative and non-discriminative patterns are selected randomly to avoid their combination into patterns of larger sizes. To simulate practical situations, for each data set generated in the above process, we add an additional 10% noise by flipping 10% of the 0's to 1's and 1's to 0's.

2) Performance of SMP and CSET on Synthetic Data Sets: For both the collections of datasets, we use a $BiggerSup$ threshold of 0.2 for CSET and a $SupMaxPair$ threshold of 0.1 for SMP. These thresholds agree with the definitions of type-I and type-II patterns for the following experiments (Section VI-A.1). The questions we want to answer in these experiments are: Which level of the itemset lattice can CSET and SMP reach when mining these synthetic datasets given the time limit of 4 hours, and correspondingly, how many of the discriminative patterns at each level can be discovered by the two algorithms?

Figures 5(a) and 5(b) display the levels that CSET and SMP reach on each of the five synthetic data sets of varying density and varying dimensionality respectively. Note that the highest level is 10, which is the size of the largest discriminative and non-discriminative patterns. Several observations can be made from Figure 5(a). First, when the density is 10%, both CSET and SMP can reach all the 10 levels. Thus, CSET can discover all the 29 type-I patterns (but none of the type-II patterns) and SMP can discover all the 21 type-II patterns (but none of the type-I patterns). Second, when the density increases to 13%, CSET only reaches level 3 and thus can only discover its 3 type-I patterns of size-2. In contrast, SMP can complete all the 10 levels and discovers all the 21 type-II patterns. Similar observation also holds for densities 0.16 and 0.19. This illustrates that even for reasonably high levels of density, SMP can discover type-II patterns with lower-support that can not be discovered by CSET, even though it can miss some type-I patterns that can be discovered by CSET. Finally, when density increases to 0.22, both SMP and CSET only reach level-2, i.e. CSET discovers its 3 type-I patterns and SMP discovers its 7 type-II patterns. This indicates that for relatively very high levels of density, both CSET and SMP can face challenges in discovering the embedded patterns that they are supposed to discover (i.e. type I patterns for CSET and type II patterns for SMP). However it should be noted that this deterioration in the performance of SMP is due to the expense of the $O(N^2)$ time complexity in the generation of level-2 candidates. Indeed, even at this density (0.22), SMP can again finish all the 10 levels in only an additional 0.5 hour (total 4.5 hrs). However, CSET is still unable to generate all the level-3 candidates even in another 4 hours (total more than 8 hours). In summary, these results show that SMP is more effective for searching for low-support discriminative patterns on dense datasets.

Similar observations can also be made from Figure 5(b). First, at the dimensionality 350, both CSET and SMP can complete all the 10 levels and discover all the patterns they are supposed to find. Second, at dimensionality 500, 2000 and 4000, CSET can only reach up to levels 6, 3 and 2 respectively, while SMP still reaches all the 10 levels. Finally, at dimensionality 6000, both SMP and CSET can only complete level 2. Again, SMP can finish all the 10 levels in another half an hour, but CSET is still generating level-3 candidates in another 4 hours. These results show that SMP is more effective for searching for low-support discriminative patterns from high-dimensional datasets.



(a) Datasets with varying density and fixed dimensionality (4000)

(b) Datasets with varying dimensionality and fixed density (0.2)

Fig. 5. Levels that can be reached by CSET and SMP in the two series of synthetic data sets (varying density and varying dimensionality).

From the above experimental results on the two collections of synthetic datasets with varying density and varying dimensionality, we demonstrated the efficacy of SMP for mining low-support discriminative patterns from dense and high-dimensional data sets. Next, we will use a real gene expression data set to study the practical utility of SMP for discovering low-support discriminative patterns.

B. Experiments on a Breast Cancer Gene Expression Data Set

In the second set of experiments, we used CSET and SMP to discover discriminative patterns from a breast cancer gene expression data set. Only closed patterns are used in these experiments. The details of this data set are provided in Section VI-B.1. We first present a global analysis of these patterns in Section VI-B.2. Subsequently, we perform an extensive statistical and biological evaluation of these patterns, the results of which are presented in Sections VI-B.3 and VI-B.4. In particular, we highlight the statistical significance and biological relevance of low-support patterns discovered by SMP but not CSET, thus illustrating the complementarity that SMP can provide to the existing approaches discussed in Section I.

1) Dataset description: A breast cancer gene expression data set [45] is used for evaluating the efficacy of discriminative pattern mining algorithms on complex, real data sets. This data set contains the expression profiles of about 25000 genes in 295 breast cancer patients, categorized into two classes corresponding to whether the patient survive the disease (0) or not (1). Using pre-processing methodologies suggested by the authors [46], we only considered 5981 genes that showed evidence of significant up- or down-regulation (at least a two-fold change), and whose expression measurements were accurate (p-value ≤ 0.01) for at least five patients. Furthermore, to make the dataset usable for binary pattern mining algorithms, each column pertaining to the expression of a single gene is split into two binary columns. Since the data has been properly normalized to eliminate between-gene variations in the scale of their expression values, we adopt a simple discretization method, as used in other studies [32], [12]: a 1 is stored in the first column if the expression of the gene is less than -0.2, while a 1 is stored in the second column if the expression of the gene is greater than 0.2. Values between -0.2 and 0.2 are not included, since genes showing an expression around 0 are not expected to be interesting, and may add substantial noise to the data set. The resulting binary data set has 11962 items and 295 transactions, with a density of 16.62%.

For this data set, discriminative pattern mining can help uncover groups of genes that are collectively associated with the progression or suppression of cancer, and our experiments are designed to evaluate the effectiveness of different algorithms for this task.

2) General analysis of the patterns discovered: We ran CSET and SMP at the lowest parameter thresholds for which they would finish in about 4 hours[5]. Only closed patterns are used in our experiments. Due to the weaker pruning of $BiggerSup$ and the resulting large number of discriminative patterns, we were forced to use relatively higher thresholds for CSET and restrict the computation to patterns of a limited size to obtain the patterns necessary for our evaluation. Table II shows that the lowest $BiggerSup$ threshold for which CSET can produce the complete results within 4 hours is 0.6. The lowest $BiggerSup$ threshold for which CSET can discover size-2 and size-3 patterns within 4 hours is 0.55. At a lower threshold of 0.4, CSET can only discover size-2 patterns before running out of time. In contrast, SMP is able to run at a much lower $SupMaxPair$ threshold of 0.18 and finds patterns of size as high as 7 in about 40 minutes. See Table III for the details of the patterns found by SMP at different thresholds. For the evaluation of pattern quality, we combine the patterns discovered by CSET at the 0.4, 0.55, and 0.6 $BiggerSup$ thresholds as the collection of all patterns that can be discovered by CSET,

[5] Some time period needed to be chosen for the experiments. The duration of four hours is, although slightly arbitrary, is generally reasonable for most data analysis operations.

$BiggerSup$ Threshold	Time (sec)	# Closed Patterns	Pattern Size(s)	Highest $NegLogP$
0.4*	617	64942	2	12.09
0.55*	1454	84840	2-3	9.65
0.6	1558	90637	2-10	8.78

TABLE II

DETAILS OF PATTERNS DISCOVERED BY CSET AT VARIOUS $BiggerSup$ THRESHOLDS. (* EXPANSION OF THE SET OF PATTERNS TO PATTERNS OF LARGER SIZES COULD NOT FINISH IN OVER 12 HOURS, AND THUS, THEIR RESULTS ARE NOT INCLUDED HERE.)

$SupMaxPair$ Threshold	Time (sec)	# Closed Patterns	Pattern Size(s)	Highest $NegLogP$
0.18	2401	45982	2-7	12.09
0.2	1187	21285	2-5	12.09
0.25	332	3007	2-4	12.09
0.3	186	283	2-3	12.09

TABLE III

DETAILS OF PATTERNS DISCOVERED BY SMP AT VARIOUS $SupMaxPair$ THRESHOLDS.

while for SMP, we only use the patterns discovered at the single $SupMaxPair$ threshold 0.18. Indeed, even with this setup that is slightly biased towards CSET, there are still high quality low-support patterns that can only be discovered by SMP, the details of which are provided later.

In addition to analyzing the characteristics of the patterns discovered by SMP and CSET, we also examined the value of DiffSup for each individual gene constituting these patterns. Specifically, Figure 6 displays the distribution of the $DiffSup$ of individual genes in the patterns discovered only by *SMP* at a SupMaxPair threshold of 0.18, but not by CSET. Among the 332 genes covered by these patterns, almost 60% (198) of the genes have $DiffSup$ lower than the 0.18. Based on the discussion of approaches that directly utilize $DiffSup$ or other measures of discriminative power for finding discriminative patterns (group C) in Section I, it can be seen that these approaches can not discover any of these genes, and thus can not discover the patterns that include them. Since one of our major foci is on algorithms that can discover patterns whose individual genes may not be discriminative, we discuss only the results of CSET and SMP, which can find such patterns, in the rest of this section.

3) Statistical Evaluation: There are various ways to evaluate the importance of discriminative patterns. We are interested in patterns that occur disproportionately between the two classes.

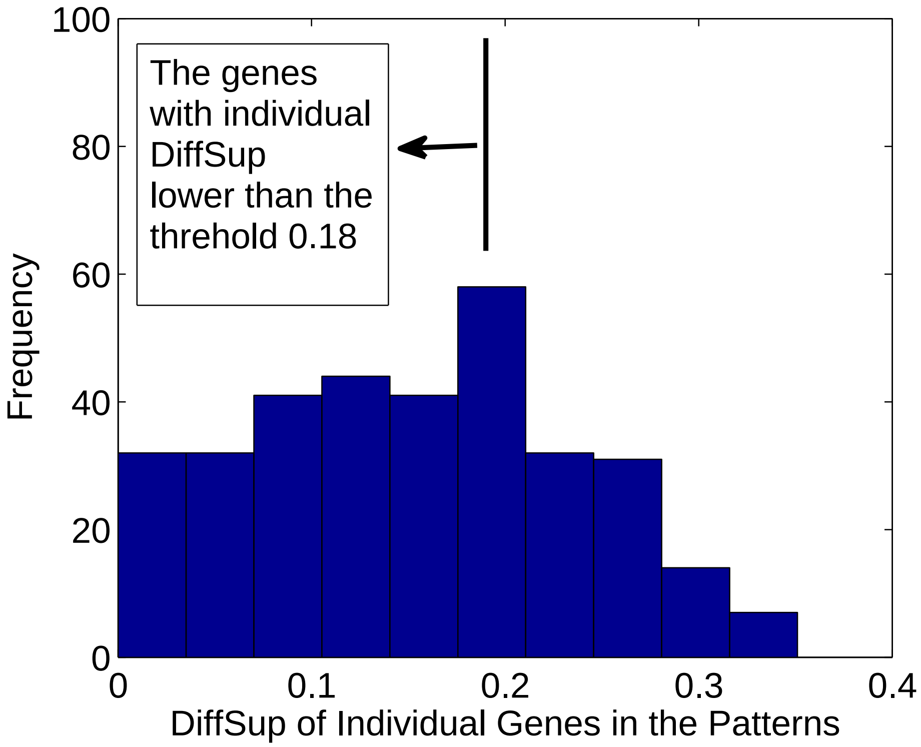
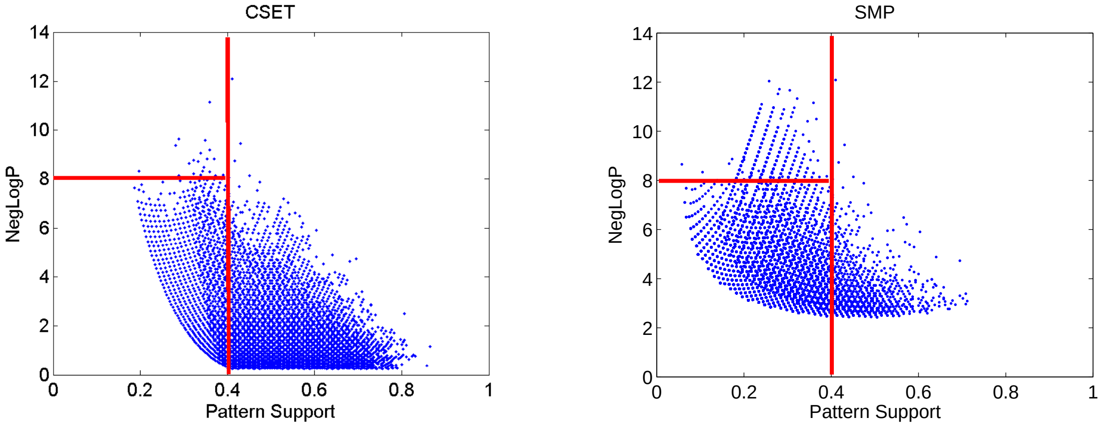


Fig. 6. Histogram of the $DiffSup$ of individual genes in the patterns discovered only by SMP, but not by CSET.

.

However, in real world data sets, particularly those with small number of instances in the two classes, even patterns that occur with similar support across classes will show some deviation from perfect balance in data sets with relatively small sample size. Thus, to ensure that the patterns found are not just a result of random fluctuation, a statistical test is commonly used to ensure that any deviation from equal support is statistically significant. In this section, we will perform this type of evaluation for the patterns from CSET and SMP.

We use the Fisher exact test [16] for this evaluation, whose result is a p-value (probability). If the p-value is below some user defined threshold, e.g., 0.05 or 0.01, then the pattern is regarded as authentic. Note that p-values are often expressed as their negative log_{10} value for convenience (the higher this $-log_{10}$ value (denoted as ***NegLogP***), the more reliable the discriminative pattern is expected to be). We will refer to this measure as ***NegLogP***. If there are multiple patterns, the $NegLogP$ threshold needs to be adjusted. By using a randomization test, as discussed below, we were able to determine that a $NegLogP$ of 8 is unlikely to arise from a random pattern. We give the technical details of this a bit later.



(a) $NegLogP$ vs. global support for CSET patterns. (b) $NegLogP$ vs. global support for SMP patterns

Fig. 7. Plot of $NegLogP$ vs. global support for patterns from CSET and SMP, where the support is relative to the whole data set.

In Figure 7, we show plots of $NegLogP$ vs. global support for the patterns discovered by both CSET and SMP. For CSET, patterns discovered by using $BiggerSup$ thresholds 0.4, 0.55, and 0.6 were combined as described in Section VI-B.2, while for SMP, a 0.18 threshold was used. Several conclusions can be drawn from this figure. First, CSET finds more patterns than SMP, particularly for patterns with higher support (the ones with support greater than 0.4). This is not surprising since SMP sacrifices completeness to find lower support patterns. Second, CSET finds many patterns with $NegLogP$ less than 2, while all the patterns discovered by SMP have $NegLogP$ higher than 2. This demonstrates the exactness of $SupMaxPair$ (Theorem 1), i.e. because $SupMaxPair$ is an lower bound of $DiffSup$, all the patterns discovered with r are $r - discriminative$. Last and the most importantly, SMP finds many patterns at low support level that are not found by CSET, especially the ones with $NegLogP$ higher than the significance threshold 8. Also, these patterns are constituted by many genes that are not covered by the patterns discovered by CSET, as will be discussed in Section VI-B.4.

We now come back to the details of how we determined a significance threshold for $NegLogP$, both for the completeness of the above discussion and to further illustrate the quality of the patterns found by SMP but not found by CSET. Because of the issues of low sample size and high-dimensionality for data sets used for problems such as biomarker discovery, many patterns may be falsely associated with the class label. This raises the multiple-hypothesis testing problem [40], [18], which are addressed by various approaches, such as Bonferroni correction [33], false discovery rate control [48] and permutation test [33], [52], [18]. Permutation tests based on row-wise, column-wise and swap randomization [18] have been used to assess the statistical significance of the results of unsupervised pattern discovery and clustering algorithms. While in the context of labeled transactions, class-label permutation tests [42], [50] are often an effective option. In this approach, a reference distribution for evaluation measures like $NegLogP$ is generated by randomly shuffling the class labels (permutations). Specifically, for each iteration, the class labels are randomly shuffled and reassigned to patients, discriminative patterns are found, and the $NegLogP$ values are computed for these patterns using the same method as for the patterns discovered with the true labels. The $NegLogP$ values from the random runs can be used to generate an empirical distribution for the $NegLogP$ values, which can be displayed as a histogram as in Figure 8. (Sometimes only the extreme (maximum) $NegLogP$ values are used as in this figure.) If a $NegLogP$ of a pattern derived from the true labels falls outside the main concentration of $NegLogP$ values from the random labels, then the $NegLogP$ very likely indicates a discriminative pattern with a ”more than random” variation from equal frequency across classes.

Figure 8 summarizes the results of such a permutation test for the dataset being used in these experiments. The right hand side shows the top 300 $NegLogP$ of the patterns discovered only by SMP but not by CSET, while the left hand side displays the maximum $NegLogP$ for each of the 1000 permutation tests where randomized labels are used for pattern mining. We observe that the $NegLogP$ values with random labels rarely exceed 8 (less than 8.72 in each of the 1000 permutation tests). Thus, we can use 8 as a relaxed threshold for significance, since only a few percent of the random patterns are above this value. The $NegLogP$ values of the top-300 patterns discovered by SMP but not by CSET with true label are much higher (all larger than 9.67). In contrast, only 34 patterns discovered by CSET have a $NegLogP$ greater than 8. This shows that SMP can discover additional statistically significant low-support patterns. In the next section, we illustrate the biological significance of these patterns and how they can be used to discover cancer-related genes.

4) Biological Relevance of Patterns based on a list of Cancer-related Genes: There are various ways to determine the biological relevance of discriminative patterns. Since the application we consider is that of discovering biomarkers for cancer, we measured the biological relevance of the patterns using a list of about 2400 human genes known to be involved in the induction, progression and suppression of various types of cancers [21]. Of these 2400 genes, 611 were

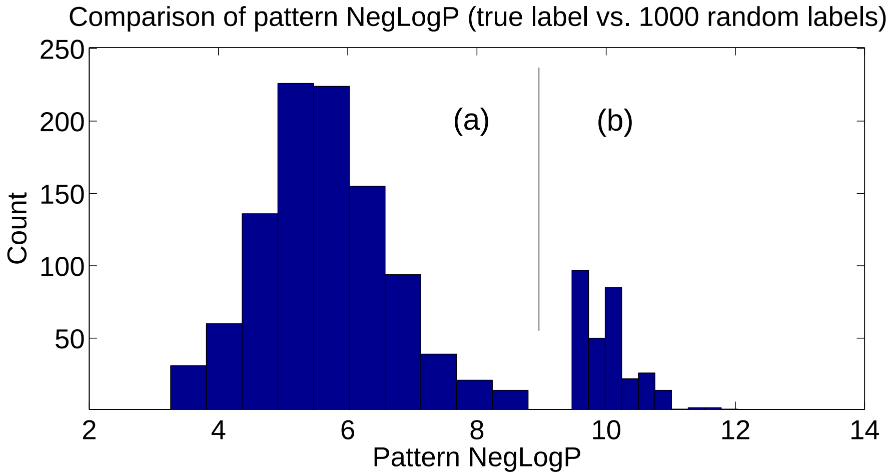


Fig. 8. Histogram of $NegLogP$ values: (a) the maximum $NegLogP$ for each of the 1000 permutation tests where randomized labels are used by SMP, (b) the top 300 $NegLogP$ values of the patterns discovered only by SMP but not by CSET

included in the set of 5981 genes in our processed gene expression data set. If the discriminative patterns found by CSET and SMP, which are just small sets of genes, tend to disproportionately contain these 610 cancer related genes as opposed to the non-cancer related genes, then this indicates that these patterns contain information that may be of significance to a biological researcher. To make this idea concrete for the purposes of evaluation, two evaluation approaches were designed.

1) **Pattern-based Biological Relevance**: For each pattern generated by CSET or SMP, we matched the genes in the pattern with the set of 611 validated cancer genes, giving us a measure of the 'precision' of the pattern. For instance, if a pattern contains 3 genes, of which 2 are found to match the list of cancer genes, then the precision of the this pattern is $2/3 = 66.67\%$. Note that if a pattern with N genes is randomly chosen from our set of 5981 genes, one would expect a precision of $[N * (611/5981)]/N = 10.2\%$.
2) **Gene Collection-based Biological Relevance**: Since patterns may overlap with each other (pattern redundancy), and do not directly show how many cancer genes can be discovered by SMP in addition to CSET, we also designed a gene collection-based evaluation methodology. Here we collect the set of genes covered by all the patterns discovered by CSET(SMP), and compare this set of genes with the set of 611 validated cancer genes just as for pattern-based evaluation. For instance, if a set of 100 patterns covers 300 genes, of

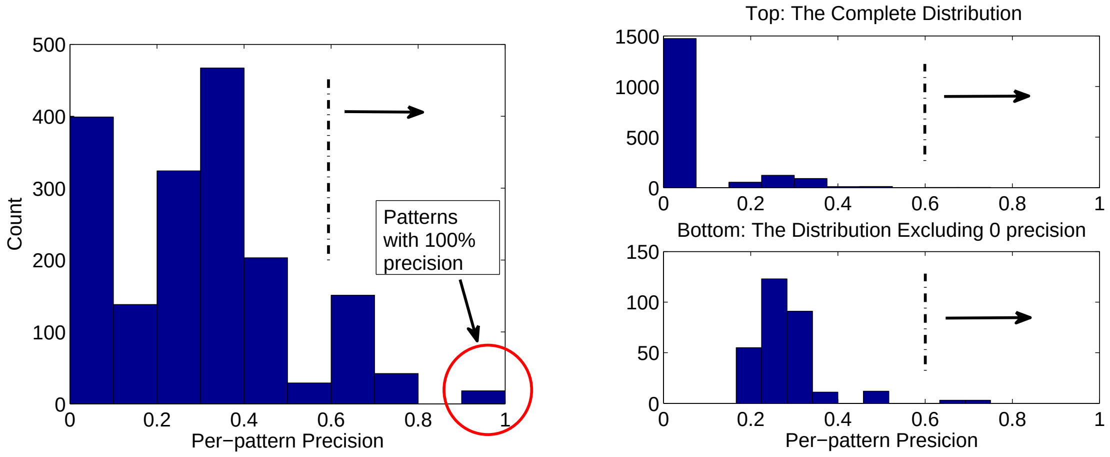
which 50 are found to match the list of cancer genes, then the precision of the set of patterns is $50/300 = 16.67\%$ and the recall is $50/611 = 8.18\%$. To compare, if we select 300 genes randomly from the 5981 genes, then the expected precision is $[300 * (611/5981)]/300 = 10.2\%$, and the expected recall is $[300 * (611/5981)]/611 = 5.02\%$.

This section details the results obtained from with these evaluation methodologies.

Brief Preview of Results: From the pattern-based biological relevance evaluation, we observed that CSET can discover patterns with good precision at relatively high support level, while SMP can further discover good quality patterns at relatively low support level, among which, there are some patterns with 100% precision with respect to the cancer gene list. From the gene collection-based biological relevance evaluation, we observed that both the techniques discovered substantially more cancer genes than expected by random chance, especially among the higher $NegLogP$ patterns. In particular, SMP was able to discover more cancer genes as compared to CSET due to its ability of discovering low-support patterns. This result further indicates the potential usefulness of recovering low-support patterns and discovering biomarkers that may be examined and utilized by the biology community. The following discussion provides additional details of these results.

Results from Pattern-based Relevance: Figure 9(a) shows the distribution of pattern-based precision of those patterns discovered only by SMP but not by CSET. For comparison, we generated a sequence of size-k patterns exactly according to the sizes of the patterns corresponding to Figure 9(a). The distribution of precision of these random patterns is shown in Figure 9(b). We can make the following observations from a comparison of Figure 9(a) and 9(b): (i) these patterns that are discovered exclusively by SMP include many that have a relatively high precision. Specifically, about 200 patterns have precisions above 0.6, among which there are 18 with a precision of 100%; (ii) the pattern-based precision of randomly generated patterns is mostly (about 1500 times) 0, and sometimes (about 300 times) fall into the range of 0.2 and 0.3, but rarely (less than 20) go beyond 0.4, and never go beyond 0.8. Interestingly, some of the SMP patterns with 100% precision play similar roles in cancer processes.

Results from Gene collection-based Relevance: To investigate how many cancer genes can be discovered using CSET and SMP, we summarized the gene collection-based evaluation results for them in Tables IV and V respectively. These tables include the number of cancer genes discovered, precision, recall, and expected recall for randomly selected group of genes of the



(a) Distribution of pattern-based precision for patterns discovered only by SMP but not by CSET

(b) Distribution of pattern-based precision of patterns generated by randomly selection of genes.

Fig. 9. Comparison of the distributions of pattern-based precision between (a) the patterns discovered by SMP but not CSET and (b) random generated patterns.

same size. Note that, the expected precision for a random collection of genes is 10.2% as calculated earlier, and thus we do not include this in these tables. The following observations can be made from these tables.

1) Both CSET and SMP usually find very precise patterns for reasonably high levels of the $NegLogP$ measure, and this precision is much higher than that expected from a set of randomly selected gene collection of the same size (10.2%). Similarly, the recall values for the genes covered by these patterns are much higher than those expected from the same type of randomly selectd gene collection, as shown by a comparison with the last column of these tables.
2) For similar values of cancer gene discovery precision, SMP generally finds more cancer genes than CSET. For instance, at a precision of about 25%, the recall of CSET is only 0.5% (3 cancer genes), while SMP has a recall 4.3% (26 cancer genes).

Note that the highlight of the second observation is not that SMP discovers more cancer genes, but that SMP can discover cancer genes from discriminative patterns with low-support in addition to the ones discovered by CSET, thus indicating the complementarity of SMP to existing approaches like CSET. Because of such complementarity, even if SMP discovered less cancer genes than CSET, SMP still complement CSET as long as additional genes are exclusively discovered by SMP. Indeed, from the specific example in the second observation, at least 23

NegLogP Threshold	# Patterns	# Genes Covered	# Cancer Genes	Pre (%)	Rec (%)	ERec (%)
12	2	3	2	66.7	0.3	0.052
11	2	3	2	66.7	0.3	0.052
10	2	3	2	66.7	0.3	0.052
9	10	12	3	25.0	0.5	0.21
8	34	31	7	22.6	1.1	0.54

TABLE IV

Precision-recall results of CSET patterns with *BiggerSup*$\geq$ 0.4 (Pre: Precision, Rec: Recall, Expected Precision for random gene collections is 10.2%, ERec: Expected Recall of Random Gene Collections with the Same Size)

NegLogP Threshold	# Patterns	# Genes Covered	# Cancer Genes	Pre (%)	Rec (%)	ERec (%)
12	2	4	2	50.0	0.3	0.067
11	6	7	3	42.9	0.5	0.12
10	200	36	11	30.6	1.8	0.60
9	541	57	17	29.8	2.8	0.95
8	1502	103	26	25.2	4.3	1.72

TABLE V

Precision-recall results of SMP patterns with $SupMaxPair\geq$ 0.18 (Pre: Precision, Rec: Recall, Expected Precision for random gene collections is 10.2%, ERec: Expected Recall of Random Gene Collections with the Same Size)

cancer genes are discovered by SMP in addition to CSET.

5) Biological Relevance of Patterns based on Biological Gene Sets: An alternative way of evaluating the biological relevance of the patterns discovered only by SMP but not by CSET is to estimate how well they capture the 5452 known biological gene sets (e.g. pathways) in the Molecular Signature database [42][6] (MSigDB). MSigDB is widely used collection of gene groups containing genes with similar biological functions. The methodology we adopt for this evaluation is one of calculating the enrichment of one pattern with these gene groups. This enrichment is measured as the probability of a random pattern of the same size having the same or better annotations by a given grne group by random chance, and the lower this probability the more enriched a pattern is with a given gene group. Specifically, for a pattern of size k

[6] Specifically, MSigDB (version 2.1, Feb 2007) contains 386 positional gene sets, 1892 curated gene sets, 837 motif gene sets, 883 computational gene sets, and 1454 annotations in Gene Ontology. http://www.broadinstitute.org/gsea/msigdb/

and a gene set of size m which share x common genes, we use the hypergeometric cumulative distribution function [7] to compute the probability that there are greater or equal to x common genes between the pattern and the gene set by random chance given that the total number of genes in the data set is N [3]. The $-log$ value of this probability can be considered as an enrichment score between a pattern and a gene set (denoted by *NegLogEnrichP*), and the larger this score, the more significant the biological relevance of the pattern. For each pattern, we use the best *NegLogEnrichP* with the 5452 gene sets as a measure of its biological relevance.

Instead of directly applying the above enrichment methodology to all the patterns that are discovered only by SMP but not by CSET, we first select a subset in which no pairs of patterns have greater than 25% overlap of genes. This selection helps reduce the effect of the redundancy between these patterns on the enrichment results. The resultant set has 37 patterns. Figure 10 shows the distribution of the best *NegLogEnrichP* values of these 37 patterns with respect to the gene sets in MSigDB. It can observed that more than half of the patterns (20) have at least two genes overlapping with one or more gene sets, and some patterns even have a *NegLogEnrichP* value as high as 8 (original p-value as low as 10^{-8}). Interestingly, some of the patterns in this collection are enriched with several gene sets that are clearly related to breast cancer such as *BREAST-DUCTAL-CARCINOMA-GENES* (*NegLogEnrichP* $= 8.02$) and *BREAST-CANCER-PROGNOSIS-NEG* (*NegLogEnrichP* $= 6.73$), as well as several gene sets that are related to general cancer-related biological processes such as the cell-growth-related gene set *IRITANI-ADPROX-LYMPH*[42] (*NegLogEnrichP* $= 6.67$) and the proliferation-related gene set *HOFFMANN-BIVSBII-BI-TABLE2*[42] (*NegLogEnrichP* $= 6.15$). These results further support the biological relevance of the patterns discovered only by SMP but not by CSET, and thus demonstrate the benefits of using SMP to search for low-support discriminative patterns in addition to existing approaches.

6) Comparison of the scalability of the algorithms: In section VI-A, we compared the effectiveness of CSET and SMP for discovering low-support patterns from synthetic datasets with varying density and dimensionality. In this part of the study, we test the scalability of CSET and SMP with varying thresholds on the real gene expression data. In addition, we also test the FPClose (FPC) [19] algorithm (plus pattern selection) as the baseline as used by other studies

[7] $p(x|k,m,N) = 1 - \sum_{i=0}^{x-1} \frac{\binom{N-i}{k-i}\binom{m}{i}}{\binom{N}{k}}$

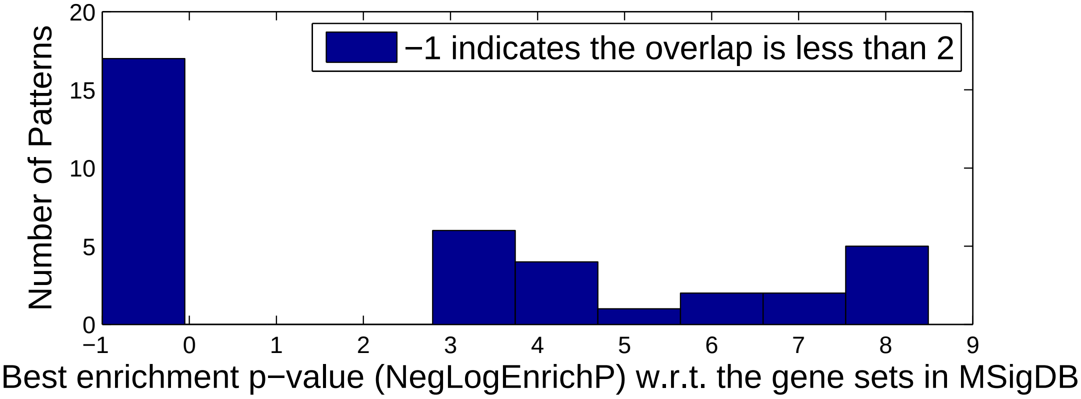


Fig. 10. Histogram of the best enrichment *NegLogEnrichP* values w.r.t. the gene sets in MSigDB, for the patterns discovered by SMP but not by CSET. An enrichment p-value is computed only if a pattern and a gene set have at least 2 genes in common.

[10], [15]. Note that, as mentioned in Section VI-B, the gene expression data set was discretized with ± 0.2 as thresholds, into a binary matrix with density 16.62% and dimension 11962, to preserve most of the information in the data. This dataset is quite dense, due to which CSET can only generate complete results at a threshold larger than 0.6. In order to obtain a more complete picture of the scalabilities of FPC, CSET and SMP, we discretized the gene expression dataset using ± 0.3 as the discretization threshold in this section, which yields a binary matrix with density 8.71%.

Figure 11 shows the results of this comparisons. The X-axis in this plot is the threshold used for discriminative pattern mining, while the Y-axis denotes the $\log_{10}$(run-time in seconds) value. Note that run-times are recorded for any algorithm only if it can produce output within four hours. The relative $minsup$ threshold used in FPC is defined on the whole dataset (both classes), while *BiggerSup* for CSET and $SupMaxPair$ for SMP take into account the support in each of the classes individually. Therefore, for a fair comparison, FPC's *minsup* is adjusted according to the size ratio of the two classes (divided by the percentage of the majority class in the whole dataset (0.74)) and then plotted together with *BiggerSup* and $SupMaxPair$.

Several observations can be made from these plots: (i) the FPC-based two-step approach can search for discriminative patterns at high support levels (above 0.55), (ii) by using *BiggerSup*, CSET is able to search at slightly lower support levels (above 0.5) compared to FPC; and for the same threshold, CSET is more efficient than FPC, and (iii) $SupMaxPair$ can explore pattern

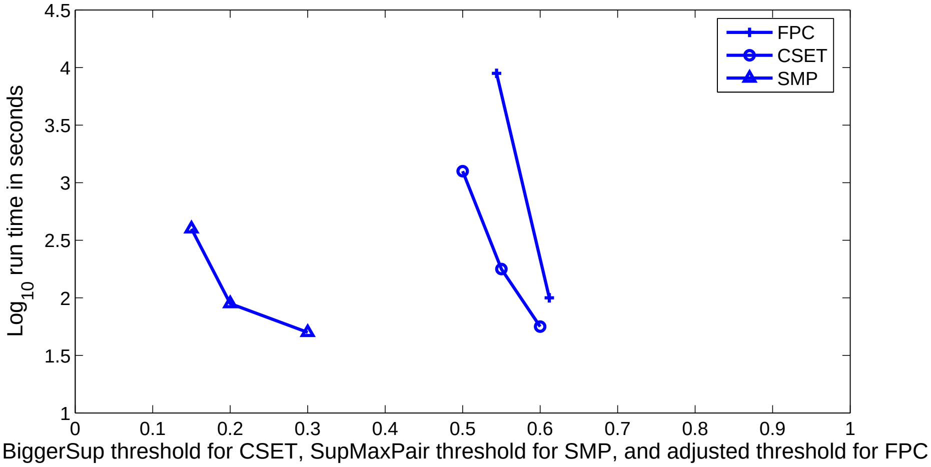


Fig. 11. Scalability of different discriminative pattern mining algorithms on the gene expression data

space with substantially lower support levels ($0.1 - 0.3$). Thus, FPC and CSET can be used to discover patterns at higher thresholds, while SMP is able to find lower support patterns missed by the other approaches.

C. Summary of Results

Based on the experimental results on both the synthetic datasets and the cancer gene expression data set presented in this section, we have demonstrated that on dense and high-dimensional data, there are patterns with relatively low support that can only be discovered by $SupMaxPair$ but not by the existing approaches. Specifically, on the cancer gene expression data set, the low-support discriminative patterns discovered only by SMP are statistically significant and biologically relevant.

We also did another set of experiments for studying how well the members of $SupMaxK$ approximate $DiffSup$ as K increases. We selected several UCI datasets [4], on which all the discriminative patterns (given a relatively low DiffSup threshold) can be discovered and used for the study. The experimental results show that: (i) $SupMax1$ generally provides very poor approximation of $DiffSup$; (ii), the approximation is improved substantially when K goes to 2, i.e. SupMaxPair; (iii) when K is increased further to 3 and 4, the computation time increases exponentially, but the approximation improves much slower compared to the improvement obtained when K goes from 1 to 2. These experimental results indicate that $SupMaxPair$ provides a good balance between the approximation of $DiffSup$ and the computational expense. The detailed results are discussed as a supplementary material (available at http://vk.cs.umn.edu/SMP/).

VII. Related work

Over the past decade, many approaches have studied discriminative pattern mining and related topics. Dong and Li [14] defined *emerging patterns (EP)* as itemsets with a sufficiently large growth rate (support ratio) between two classes. A two-step algorithm is proposed to discover EPs, which first finds frequent itemsets with the Max-Miner algorithm [6] for each of the two classes, and then compares these itemsets to find EPs. Emerging pattern were the first formulation of discriminative patterns and have been extended further to several special cases such as jumping emerging patterns [26] and minimal emerging patterns[31], [27]. Here, the discriminative power of a pattern is measured with support ratio [14], or simply with the two supports of the pattern in the two classes and two corresponding thresholds [31]. As discussed in [5], these emerging pattern mining algorithms must mine the data multiple times given a certain threshold for support ratio (or two thresholds for the two supports). In [5], a new formulation of discriminative patterns, contrast sets (CSETs), is proposed along with an algorithm to mine them. CSET is the first technique that formulates discriminative pattern mining within an Apriori-like framework [2], [6], in which different pruning measures can be used to perform a systematic search on the itemset lattice [2]. In [51], contrast set mining is shown to be a special case of a more general task, namely rule learning, where a contrast set can be considered as an antecedent of a rule whose consequent is a group. Notably, CSET has also be used in some biomedical applications [25]. The upper bounds of statistical discriminative measures have also been studied for discriminative pattern mining e.g. information gain [9], $\chi^2 - test$ [5] and several others [34].

Next, we also briefly discuss other research work related to discriminative pattern mining, although they are not the focus of the paper. Many existing approaches have studied the use of frequent patterns in classification. Associative classifiers [29], [28], [55], [11], [49] are a series of approaches that focus on the mining of high-support, high-confidence rules that can be used in a a rule-based classifier. Cheng et. al. [9] recently conducted a systematic evaluation of the utility of frequent patterns in classification. Several pattern-based classification frameworks have also been proposed, in which a small number of discriminative patterns are selected, which can achieve comparable classification accuracy with respect to the whole set of discriminative patterns [10], [15], [54], [30]. Discriminative pattern mining from multiple classes has been studied in [5], [27], [25], while mining complex discriminative patterns has been studied in [31]. Although traditional pattern summarization approaches [20] can be adopted to control the redundancy among discriminative patterns, closeness and redundancy are specially studied for in the context of discriminative patterns respectively in [17] and [41].

VIII. Conclusions

In this paper, we addressed the necessity of trading off the completeness of discriminative pattern discovery, with the ability to discover low-support discriminative patterns from dense and high-dimensional data within an acceptable amount of time. For this, we proposed a family of anti-monotonic measures of discriminative power named $SupMaxK$ that conceptually organize the set of discriminative patterns into nested layers of subsets, and are progressively more complete in their coverage, but require increasingly more computation for their discovery. Given the same and fixed amount of time, the $SupMaxK$ family provides a tradeoff between the ability to search for low-support discriminative patterns and the coverage of the space of valid discriminative patterns for the corresponding threshold. In particular, $SupMaxK$ with K = 2 named $SupMaxPair$, is a special member of this family that is suitable for dense and high-dimensional data. We designed a framework, named SMP, which uses $SupMaxPair$ for discovering discriminative patterns from dense and high-dimensional data. A variety of experiments on both synthetic datasets and a breast cancer gene expression dataset demonstrated that there are patterns with relatively low support that can be discovered using SMP but not by the existing approaches. In particular, the low-support discriminative patterns discovered only by SMP from the gene expression dataset are statistically significant and biologically relevant. In summary, SMP can complement existing algorithms for discovering discriminative patterns by finding patterns with relatively low support from dense and high-dimensional data sets that other approaches fail to discover within an acceptable amount of time. Thus, in practice, it is recommended that CSET and other existing approaches should be used to discover medium-to-high support patterns from such data sets within an acceptable amount of time, and then SMP could be used to further discover low-support discriminative patterns that existing approaches may not discover.

Our work can be extended in several directions. As discussed in Section IV-D, the members of $SupMaxK$ induce a hierarchy of subsets of the complete set of discriminative patterns. This hierarchy motivates further research that focuses on the mining of discriminative patterns from the other layers that are not covered by $SupMaxPair$. It is also interesting to study the quality of the discriminative patterns in the different layers of this hierarchy, which may provide insights into different priorities for discriminative pattern mining from these layers. Note that, the use of measures from the $SupMaxK$ family is only one possible method for trading off the completeness of pattern discovery with the ability to discover low-support discriminative patterns from high-dimensional data. Indeed, other approaches that adopt a different strategy for handling this tradeoff are also possible and should be studied. Also, most existing discriminative pattern mining algorithms (as well as SMP) are designed for binary data, and have to rely on discretization for continuous data. It will be useful to design approaches that can directly handle continuous data for discriminative pattern mining, as has been done for discovering patterns in an unsupervised manner[36].

IX. Supplementary Material: Experiments to show how well *SupMaxK* estimates *DiffSup*

In this set of experiments, we study how the members of *SupMaxK* approximate *DiffSup* as K increases. There are two approaches for this purpose, i.e. analytical and empirical analysis. In an analytical approach, some assumptions need to be made such as that the data comes from independence model. Such assumptions generally do not hold for real datasets. Therefore, we selected several real datasets from UCI Data Repository [4] and designed an empirical study on the approximation of *DiffSup* by the members of *SupMaxK*. The datasets we selected are mushroom, hypo, hepatic and sonar, which have relatively low density or low dimensionality, so that a low *DiffSup* threshold (0.1) can be used to discover the complete set of discriminative patterns for a comprehensive study on the approximation.

Given a dataset, for each discriminative pattern of size no less than N, we compute the value of its *SupMax1*, *SupMax2*, ... and *SupMax(N-1)*, and compare this sequence of values with its *DiffSup*, from which we can see how *SupMaxK* approximates *DiffSup* with increasing value of K. The results on these datasets are displayed in Figures 12. Several observations can be made:

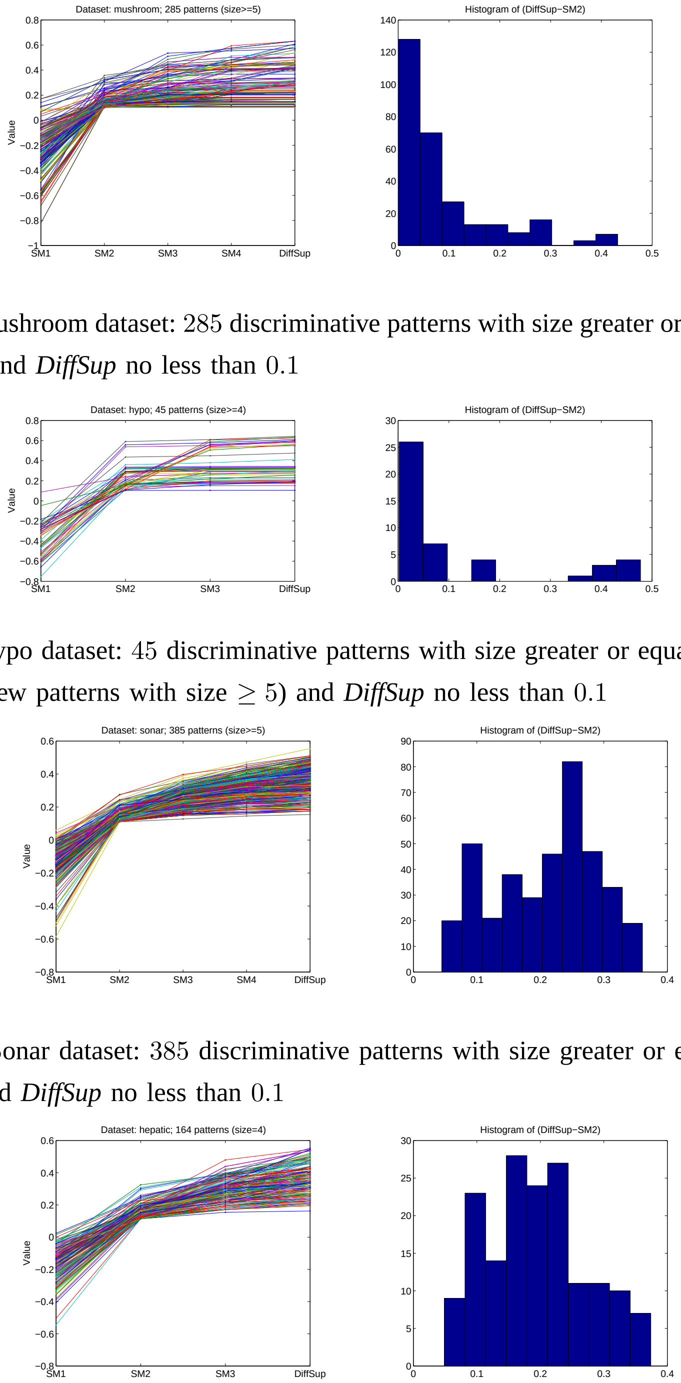
- Firstly, as K increases, *SupMaxK* provides a closer and closer approximation of *DiffSup*. Specifically in the left subfigures, all the patterns have non-decreasing *SupMaxK* values (shown by the non-decreasing curves). This observation is guaranteed by Lemma 3 and Theorem 1.

- Secondly, *SupMax1* generally provides very poor approximation of DiffSup. Specifically, although all the patterns discovered from the four datasets have *DiffSup* no less than 0.1, most of them have negative *SupMax1* values.

- Thirdly, when K goes from 1 to 2, i.e. *SupMaxPair*, the approximation is improved substantially (shown by the jump of value from $SM1$ to $SM2$). With this improvement, for many discriminative patterns, *SupMaxPair* ($K = 2$) provides a reasonably good approximation of *DiffSup*. Take the mushroom dataset as an example, for 200 of the total 285 patterns (70.2%), *SupMaxPair* has difference less than 0.1 from *DiffSup*. There are also many patterns whose *SupMaxPair* values have differences less than 0.1 from *DiffSup* in the other three datasets: hypo (about (70%)), sonar (about (20%)) and hepatic (about (20%)).

- Finally, when K is increased further to 3 and 4, the computation time increases exponentially, but the approximation improves relatively much less compared to the improvement obtained when K goes from 1 to 2. However, it is worthnoting that the differences between *SupMaxPair* and *DiffSup* can also be large (ranging from 0.1 to 0.4) for many discriminative patterns on all the four datasets, e.g. 30% in the mushroom dataset, about 80% in the hepatic and sonar datasets. For these discriminative patterns, *SupMaxK* with larger K (≥ 3) is necessary to provide sufficiently close approximation of *DiffSup*.

These experimental results indicate that *SupMaixPair* provides a good balance between the approximation of *DiffSup* and the computational expense. However, we present the details of this study in the supplementary material rather than in the main paper due to the space limits (we are alrady at the maximal 36 pages allowed), and because the highlight of *SupMaxPair* is not its accurate approximation of *DiffSup* but the combination of the following three advantages: (i) it is effective for pruning non-discriminative patterns as a lower bound of *DiffSup*, compared to *BiggerSup* (an upper bounds of *DiffSup*) (ii) it is a tighter lower bound for DiffSup compared to *SupMax1* (theoretically guaranteed by Lemma 3 and Theorem 1, and also shown in Figure 12) and (iii) it is the only one, among the members of *SupMaxK* ($K \geq 2$), that is feasible for handling high dimensional datasets. These advantages enable *SupMaxPair* for discovering additional low-support discriminative patterns from dense and high-dimensional dataset when existing techniques fail to, as extensively discussed in Sections I and IV and demonstrated in Sections VI-A and VI-B.



(a) Mushroom dataset: 285 discriminative patterns with size greater or equal to 5 and *DiffSup* no less than 0.1

(b) Hypo dataset: 45 discriminative patterns with size greater or equal to 4 (too few patterns with size ≥ 5) and *DiffSup* no less than 0.1

(c) Sonar dataset: 385 discriminative patterns with size greater or equal to 5 and *DiffSup* no less than 0.1

(d) Hepatic dataset: 164 discriminative patterns with size greater or equal to 4 (too few patterns with size ≥ 5) and *DiffSup* no less than 0.1

Fig. 12. The approximation of *DiffSup* by the members of *SupMaxK* with increasing value of K on the three UCI data sets. In the left subfigures, the sequence of values for each pattern (*SupMax1*, *SupMax2*, *SupMax3*, *SupMax4* and *DiffSup*) are plotted as a curve. The right subfigures are the distribution of the difference between *DiffSup* and *SupMax2*, the value of which measures how close *SupMax2* approximate *DiffSup*

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for the constructive comments. This work was supported by NSF grants #IIS0916439, #CRI-0551551, a University of Minnesota Rochester Biomedical Informatics and Computational Biology Program Traineeship Award. Access to computing facilities was provided by the Minnesota Supercomputing Institute.

REFERENCES

[1] Mental Health Services Administration (2006). The role of biomarkers in the treatment of alcohol use disorders. *Substance Abuse Treatment Advisory*, 5(4):4206–4223.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. VLDB*, pages 487–499, 1994.

[3] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[4] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

[5] S. Bay and M. Pazzani. Detecting group differences: Mining contrast sets. *DMKD*, 5(3):213–246, 2001.

[6] R. J. Bayardo. Efficiently mining long patterns from databases. In *Proc. SIGMOD*, pages 85–93, 1998.

[7] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proc. SIGMOD*, pages 265–276, 1997.

[8] C. Carlson et al. Mapping complex disease loci in whole-genome association studies. *Nature*, 429(6990):446–452, 2004.

[9] H. Cheng, X. Yan, J. Han, and C.-W. Hsu. Discriminative frequent pattern analysis for effective classification. In *Proc. ICDE*, pages 716–725, 2007.

[10] H. Cheng, X. Yan, J. Han, and P. Yu. Direct discriminative pattern mining for effective classification. In *Proc. ICDE*, pages 169–178, 2008.

[11] G. Cong, K. Tan, A. Tung, and X. Xu. Mining top-k covering rule groups for gene expression data. In *Proc. SIGMOD*, pages 670–681, 2005.

[12] C. Creighton and S. Hanash. Mining gene expression databases for association rules. *Bioinformatics*, 19(1):79–86, 2003.

[13] M. Deshpande, M. Kuramochi, N. Wale, and G. Karypis. Frequent sub-structure based approaches for classifying chemical compounds. *IEEE TKDE*, 17(8):1036–1050, 2005.

[14] G. Dong and J. Li. Efficient mining of emerging paterns: Discovering trends and differences. In *Proc. SIGKDD*, pages 43–52, 1999.

[15] W. Fan, K. Zhang, H. Cheng, J. Gao, X. Yan, J. Han, P. S. Yu, and O. Verscheure. Direct mining of discriminative and essential graphical and itemset features via model-based search tree. In *Proc. SIGKDD*, pages 230–238, 2008.

[16] R. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, pages 87–94, 1922.

[17] G. Garriga, P. Kralj, and N. Lavrač. Closed sets for labeled data. *JMLR*, 9:559–580, 2008.

[18] A. Gionis et al. Assessing data mining results via swap randomization. *ACM TKDD*, 1(3), 2007.

[19] G. Grahne and J. Zhu. Efficiently using prefix-trees in mining frequent itemsets. In *Workshop on Frequent Itemset Mining Implementations*, 2003.

[20] J. Han et al. Frequent pattern mining: current status and future directions. *DMKD*, 15:55–86, 2007.

[21] M. E. Higgins, M. Claremont, J. E. Major, C. Sander, and A. E. Lash. CancerGenes: a gene selection resource for cancer genome projects. *Nucl. Acids Res.*, 35(supplement 1):D721–726, 2007.

[22] T. Hwang, H. Sicotte, Z. Tian, B. Wu, J. Kocher, D. Wigle, V. Kumar, and R. Kuang. Robust and efficient identification of biomarkers by classifying features on graphs. *Bioinformatics*, 24(18):2023–2029, 2008.

[23] S. Jaroszewicz and D. A. Simovici. Pruning redundant association rules using maximum entropy principle. In *Proc. PAKDD*, pages 135–147, May 2002.

[24] P. Kralj, N. Lavrac, D. Gamberger, and A. Krstacic. Contrast set mining for distinguishing between similar diseases. In *AI in Medicine in Europe*, pages 109–118, 2007.

[25] P. Kralj Novak, N. Lavrač, D. Gamberger, and A. Krstačić. CSM-SD: Methodology for contrast set mining through subgroup discovery. *Journal of Biomedical Informatics*, 42(1):113–122, 2009.

[26] J. Li, G. Dong, and K. Ramamohanarao. Making use of the most expressive jumping emerging patterns for classification. *KAIS*, 3(2):131–145, 2001.

[27] J. Li, G. Liu, and L. Wong. Mining statistically important equivalence classes and delta-discriminative emerging patterns. In *Proc. SIGKDD*, pages 430–439. ACM New York, NY, USA, 2007.

[28] W. Li, J. Han, and J. Pei. CMAR: Accurate and efficient classification based on multiple class-association rules. In *Proc. ICDM*, pages 369–376, 2001.

[29] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *Proc. SIGKDD*, pages 80–86, 2001.

[30] D. Lo, H. Cheng, J. Han, S. Khoo, and C. Sun. Classification of software behaviors for failure detection: a discriminative pattern mining approach. In *Proc. SIGKDD*, pages 557–566, 2009.

[31] E. Loekito and J. Bailey. Fast mining of high dimensional expressive contrast patterns using zero-suppressed binary decision diagrams. In *Proc. SIGKDD*, pages 307–316. ACM New York, NY, USA, 2006.

[32] T. McIntosh and S. Chawla. High confidence rule mining for microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(4):611–623, 2007.

[33] R. Miller. Simultaneous Statistical Inference. *Springer-Verlag Inc., New York, NY*, 1981.

[34] S. Morishita and J. Sese. Transversing itemset lattices with statistical metric pruning. In *Proc. PODS*, pages 226–236. ACM New York, NY, USA, 2000.

[35] P. Novak, N. Lavrac, and G. Webb. Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. *JMLR*, 10:377–403, 2009.

[36] G. Pandey, G. Atluri, M. Steinbach, C. L. Myers, and V. Kumar. An Association Analysis Approach to Biclustering. In *Proc. SIGKDD*, pages 677–686, 2009.

[37] N. Pasquier, Y. Bastide, R. Taouil, , and L. Lakhal. Discovering frequent closed itemsets for association rules. In *Proc. ICDT*, pages 398–416, 1999.

[38] E. Segal, N. Friedman, N. Kaminski, A. Regev, and D. Koller. From signatures to models: understanding cancer using microarrays. *Nature Genetics*, pages S38–S45, 2005.

[39] D. Segre et al. Modular epistasis in yeast metabolism. *Nature Genetics*, 37:77–83, 2004.

[40] J. Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46(1):561–584, 1995.

[41] A. Soulet et al. Condensed representation of emerging patterns. In *Proc. PAKDD*, pages 127–132, 2004.

[42] A. Subramanian, P. Tamayo, V. Mootha, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005.

[43] P.-N. Tan, M. Steinbach, and V. Kumar. Introduction to data mining. *Addison-Wesley*, 2005.

[44] N. Tatti. Maximum entropy based significance of itemsets. *Knowledge and Information Systems*, 17(1), Oct 2008.

[45] van de Vijver et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347:1999–2009, 2002.

[46] L. J. van ’t Veer et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[47] M. van Vliet, C. Klijn, L. Wessels, and M. Reinders. Module-based outcome prediction using breast cancer compendia. *PLoS ONE*, 2(10):1047, 2007.

[48] K. Verhoeven et al. Implementing false discovery rate control: increasing your power. *Oikos*, 108(3):643–647, 2005.

[49] J. Wang and G. Karypis. HARMONY:efficiently mining the best rules for classification. In *Proc. SDM*, page 205, 2005.

[50] K. Wang et al. Pathway-based approaches for analysis of genomewide association studies. *AJHG*, 81(6):1278–1283, 2007.

[51] G. I. Webb et al. On detecting differences between groups. *Proc. SIGKDD*, pages 256–265, 2003.

[52] P. Westfall and S. Young. P value adjustments for multiple tests in multivariate binomial models. *JASA*, page 780, 1989.

[53] H. Xiong, P. Tan, and V. Kumar. Hyperclique pattern discovery. *DMKD*, 13(2):219–242, 2006.

[54] X. Yan, H. Cheng, et al. Mining significant graph patterns by leap search. In *Proc. SIGMOD*, pages 433–444, 2008.

[55] X. Yin and J. Han. CPAR: Classification based on predictive association rules. In *Proc. SDM*, pages 331–335, 2003.

[56] N. Yosef, Z. Yakhini, A. Tsalenko, V. Kristensen, A. Borresen-Dale, E. Ruppin, and R. Sharan. A supervised approach for identifying discriminating genotype patterns and its application to breast cancer data. *Bioinformatics*, 23(2):91–98, 2007.