Bridging Causal Relevance and Pattern Discriminability: Mining Emerging Patterns from High-Dimensional Data*

Kui Yu[1], Wei Ding[2], Hao Wang[1], and Xindong Wu ✉ [1,3]

Abstract—It is a nontrivial task to build an accurate Emerging Pattern (EP) classifier from high-dimensional data because we inevitably face two challenges (1) how to efficiently extract a minimal set of strongly predictive EPs from an explosive number of candidate patterns, and (2) how to handle the highly sensitive choice of the minimal support threshold. In order to address these two challenges, we bridge causal relevance and EP discriminability (the predictive ability of emerging patterns) to facilitate EP mining and propose a new framework of mining EPs from high-dimensional data. In this framework, we study the relationships between causal relevance in a causal Bayesian network and EP discriminability in EP mining, and then reduce the pattern space of EP mining to direct causes and direct effects, or the Markov blanket of the class attribute in a causal Bayesian network. The proposed framework is instantiated by two EPs-based classifiers, CE-EP and MB-EP, where CE stands for direct ***C***auses and direct ***E***ffects, and MB for ***M***arkov ***B***lanket. Extensive experiments on a broad range of datasets validate the effectiveness of the CE-EP and MB-EP classifiers against other well-established methods, in terms of predictive accuracy, pattern numbers, running time, and sensitivity analysis.

Index Terms—Emerging Patterns, Causal Bayesian Networks, Causal Relevance, EP Discriminability

1 INTRODUCTION

Association rule mining seeks to find association patterns that meet predefined minimum support and confidence constraints from a given dataset [7, 23]. This problem is usually divided into two steps. The first is to find frequent itemsets whose supports exceed a predefined minimum support threshold; and the second is to generate association rules from those frequent itemsets with the constraint of minimal confidence [32-33]. Associative classification integrates association rule mining and classification [22, 35]. In associative classification, the consequent of an association rule is a class label, and the classifier is constructed using a set of association rules. This classifier is expected to produce accurate classifications and yield an interpretable model [3, 16]. Liu et al. [22] introduced CBA (Classification Based on Associations), the first associative classifier.

An illustrating example of association rules for classification is given in Table 1 using the Balloon dataset from the UCI machine learning repository [4], with the class attribute, *inflated* (T (true), F

A shorter, preliminary version of this paper with the title "Causal Associative Classification" was published in the Proceedings of the 11th IEEE International Conference on Data Mining (ICDM'11), pp. 914-923. Compared with the conference version of this paper, this journal version has been completely re-written with the following new materials: Tables 1 and 2 in Section 1; Section 2.2; Section 3.1; a new theoretical analysis of causal relevance in causal Bayesian networks and EP discriminability in EP mining in Section 4; the number of selected datasets has been increased from 20 to 36 in Section 5; new Sections 5.3 and 5.5 to have a detailed analysis on comparisons of the numbers of selected patterns and on sensitivity analysis on both predictive accuracy and numbers of selected patterns under seven minimum support thresholds, respectively; and new Sections 5.2.2, 5.6 and 5.7.

[1] *Department of Computer Science, Hefei University of Technology, Hefei, 230009, China. E-mail: ykui713@ gmail.com, jsjwangh@hfut.edu.cn.*
[2] *Department of Computer Science, University of Massachusetts Boston, Boston, 02125, USA. E-mail: ding@cs.umb.edu.*
[3] *Department of Computer Science, University of Vermont, Burlington, 05405, USA. E-mail: xwu@cs.uvm.edu.(✉ corresponding author)*

(false)), and 4 features: *color* (yellow, purple), *size* (large, small), *act* (stretch, dip) and *age* (adult, child). The dataset consists of 20 samples as listed in Table 3 of Section 3.1.

If we set the minimum support threshold to 0.2 and the minimum confidence threshold to 0.8, the top five association rules mined from this dataset for classification are shown in Table 1.

TABLE 1 EXAMPLES OF CLASSIFICATION ASSOCIATION RULES

ID	Association rule
1	act=stretch==>inflated=T
2	age=adult==>inflated=T
3	act=dip & age=child==>inflated=F
4	color=yellow & act=stretch ==>inflated=T
5	color=purple & act=stretch ==>inflated=T

Later, Dong and Li proposed a new type of association patterns named ***E***merging ***P***atterns (EPs for short) whose support values change significantly from one class to another [9]. Different from association rule mining, a data set is divided into several subsets by their class labels in EP mining. The ratio of the support of an itemset in one class and that of this itemset in a contrasting class is measured using the growth rate. Those patterns whose growth rates satisfy a predefined minimum threshold are called EPs. Hence, EPs represent strong contrasts between different classes of data. For example, to mine EPs from the Balloon dataset, this dataset is divided into two classes: *inflated* =T and i*nflated* =F before mining, then the EPs of each class are mined from the corresponding class data, respectively, as shown in Table 2, under the minimum support threshold of 0.2 and a growth rate greater than 1.

TABLE 2 EXAMPLES OF EPS MINED FROM BALLOON DATASET

ID	EPs of the class inflated=T	EPs of the class inflated=F
1	act=stretch	act=dip
2	age=adult	age=child
3	-	(act=dip) & (age=child)

From the example above, we can see that EPs give more concise and understandable patterns than association rules. Moreover, the presence of EPs gives evidence about which class the object should belong to. Thus, the discovery process of EPs prefers classification. Dong et al. proposed the first EPs-based classifier, called CAEP (Classification by Aggregating Emerging Patterns)[10]. EPs-based classification has shown to be a powerful method for constructing accurate classifiers, even for imbalanced data [10-11, 13]. This paper focuses on mining EPs for classification.

Most associative classifiers are constructed in two steps: generating frequent patterns satisfying minimum support and confidence constraints, and then making predictions based on the selected patterns. Although many pruning strategies have been proposed, an explosive number of rules can still be discovered from high dimensional and dense data even using a rather high minimum support threshold. The large number of candidate rules makes it difficult to store, retrieve, prune, and sort them efficiently for classification. Furthermore, they hamper the understanding of the final classifiers, and even

lead to overfitting. Hence how to select a suitable minimum support threshold is not only a challenging problem, but also the key to control the performance of associative classifiers. A small support threshold could generate a large number of rules while a large value might prune many predictive rules and cause serious accuracy degradation.

As a special type of association mining, mining EPs from high-dimensional data also encounters the above challenging problem, especially with the advent of the emerging datasets with tens of thousands of features in many real-world applications, such as image processing, gene expression data, text data, etc. Thus, to effectively mine EPs from high-dimensional data, two challenging research issues need to be further explored:

(1) How to efficiently mine a minimal set of strongly predictive EPs from high-dimensional data; and

(2) How to deal with the highly sensitive choice of the minimal support threshold.

To battle these challenges, we propose a new framework for mining EPs from high-dimensional data by bridging causal relevance in causal Bayesian networks and EP discriminability (the predictive ability of EPs) in EP mining. More specifically, the causal relevance of a target node in causal Bayesian networks with respect to other nodes is divided into three categories, irrelevant nodes, Markov blanket (direct causes, direct effects and direct causes of the direct effects of the target node), and redundant nodes while the pattern space in EP mining with respect to EP discriminability is classified as non-EPs, strongly predictive EPs, and redundant EPs. Through studying the relationships between causal relevance in a causal Bayesian network and EP discriminability in EP mining, we bridge causal relevance and EP discriminability to reduce the pattern space in EP mining to the direct ***C***auses and direct ***E***ffects (CE), or the ***M***arkov ***B***lanket (MB) of the class attribute in causal Bayesian networks to facilitate EP mining in an innovative framework and mine EPs from high-dimensional data.

The main contributions of this paper are as follows:

(1) The paper gives a theoretical analysis of the relationships between causal relevance in causal Bayesian networks and EP discriminability in EP mining. With this theoretical framework, the pattern space in EP mining is reduced to the space of CE or MB of the class attribute in a causal Bayesian network instead of the combinations of all features, which greatly reduces computational cost and resource demand in the stage of EP mining.

(2) By bridging causal relevance with EP discriminability, mining EPs from the space of CE or MB of the class attribute in a causal Bayesian network, naturally endows EPs with strongly predictive ability, since the causal factors of a variable give a natural interpretation of the events occurring in real-world applications. Most importantly, in a causal Bayesian network, the CE or the MB of a target node is unique and minimal, and hence, our framework has a good chance to generate a minimum set of EPs.

(3) With the above innovative framework, two new EPs-based classifiers, CE-EP and MB-EP, are proposed. Extensive experiments on a broad range of datasets, including 24 UCI datasets and 12 very high-dimensional datasets, validate the effectiveness of the proposed approaches against other well-established methods, in terms of predictive accuracy, pattern numbers, and running time.

(4) The experiments of sensitivity analysis on seven minimum support thresholds demonstrate that the CE-EP and MB-EP classifiers not only can efficiently and effectively handle very high-dimensional data, but also are insensitive to the minimum support thresholds. Moreover, our experiments discover that the EPs-based classifiers are less sensitive to the minimum support threshold than the associative classifiers, and the choice of a suitable minimum support threshold is the key to control CBA and CMAR while it is not crucial to CAEP, especially to both CE-EP and MB-EP. Finally, our study of impact of the minimal growth-rate threshold illustrates that both CAEP and CE-EP classifiers are also less sensitive to the minimal growth-rate threshold.

The reminder of the paper is structured as follows. Section 2 reviews previous work. Section 3 provides the backgrounds on emerging patterns and causal Bayesian networks, and Section 4 bridges causal relevance with EP discriminability, and then presents a framework for mining EPs from high-dimensional data. Experimental results are reported in Section 5, and we conclude in Section 6.

2 PREVIOUS WORK

2.1 EPs-based Classifiers

Associative classification integrates association rule discovery and classification into a prediction model. Successful algorithms of associative classifiers include CBA [22], CMAR [19] and CPAR [36]. CBA (Classification Based on Association) uses an Apriori-like algorithm to generate a single rule-set and ranks the rules according to their confidence/support values. Then CBA adopts "one matching pattern determines the class of an instance" approach to select the best rule to be applied to each test instance. Based on CBA, Li et al. introduced CMAR (Classification based on Multiple-class Association Rule) that generates classification association rules through a FP-tree and uses multiple rules to perform the classification, while CPAR (Classification based on Predictive Association Rule) combines the advantages of both associative classification and traditional rule-based classification. Instead of generating a large number of candidate rules as in associative classification, CPAR adopts a greedy algorithm to generate rules directly from the training data.

Dong and Li introduced Emerging Patterns (EPs) to represent strong contrasts between different classes of data [9]. An emerging pattern is a multivariate pattern whose support value increases sharply from a background dataset to a target dataset. Compared to association rules, EPs capture emerging trends in time-stamped datasets, or useful contrasts between data classes [9, 28]. In addition, Jumping

Emerging Patterns (JEPs, as defined in Section 3.1) is a special type of EPs whose supports increase from zero in a background dataset to non-zero in a target dataset [17]. Like other patterns or rules composed of conjunctive combinations of attributes and values, EPs can be easily understood and used directly in a wide range of applications, such as predicting diseases [20], failure detection [24], and discovering knowledge in gene expression data [5, 12, 21].

EPs represent strong contrasts between different classes of data, and the presence of EPs in a query object gives some evidence about which class the object should belong to. Therefore, EPs have shown very successful results on constructing accurate and robust classifiers. In comparison with associative classifiers based on association rules, EPs-based classifiers use the aggregation of the discriminating power of the set of matching EPs to classify an instance. Dong et al. proposed the first EPs-based classifier, called CAEP (Classification by Aggregating Emerging Patterns)[10]. In fact, both CMAR and CPAR have adopted the idea of CAEP by using multiple rules instead of one rule to classify an instance. CAEP first discovers all the EPs from the training data for each class. When a new test instance is classified by aggregating the differentiating power of a set of EPs that apply, a score is computed for each class, and this test instance is classified to the class with the highest score. Based on CAEP, Li et al. proposed a JEP-classifier which is distinct from the CAEP classifier [17]. The JEP-classifier uses JEPs exclusively because JEPs discriminate between different classes more strongly than any other type of EPs. Since discovery of all EPs from the training data is time consuming, Li et al. [18] presented a lazy EPs-based classifier, called DeEPs, to improve the efficiency and accuracy of CAEP and JEP-classifier. Whenever a new test instance is considered, DeEPs uses it as a filter to remove irrelevant feature values in order to reduce the search space. Since an EP mining process of DeEPs is instance-based, all the training data has to be stored for re-learning during the entire classification process. Fan and Ramamohanarao proposed a robust EP-classifier named SJEP-classifier, exclusively using a strong JEP [11]. A strong JEP from the class C_1 to the class C_2 satisfies two conditions: (1)the support of itemset X is zero in C_1 but non-zero in C_2 and satisfies a minimal support threshold in C_2, and (2) any proper subset of X does not satisfy condition (1). The SJEP-classifier integrates the CP-tree data structure into the EP classifier, which uses far fewer JEPs than a JEP-classifier yet gets higher predictive accuracy than existing EPs-based classifiers.

Due to the limitation of the existing EP mining techniques, existing EP-based classifiers could not effectively handle datasets with more than sixty dimensions without prior feature set reduction using desktop computers of 2006 [25]. It is still a challenging research issue to build an accurate EPs-based classifier from a high-dimensional dataset. In this study, we bridge causal relevance in causal Bayesian networks and EP discriminability in EP mining to help construct accurate EPs-based classifiers from high-dimensional data.

2.2 Learning Causal Bayesian Networks from Data

A causal Bayesian network is a Bayesian network in which each directed edge is described as a direct causal influence imposed on a child node by its parent nodes. Since structure learning of causal Bayesian networks in observational data is essentially the same as structure learning of Bayesian networks, learning Bayesian networks is one of the most common methods to explore causal relationships in the observed data [29, 31]. Structure learning methods of Bayesian networks include global and local learning approaches. A global learning approach attempts to uncover a complete Bayesian network over all model features, but it can only deal with no more than 300 features [6, 8].

A local learning approach without learning a complete Bayesian network has been considered as an effective means to handle hundreds of thousands of features [2, 34]. The local learning focuses on two specific tasks: (a) identification of features that are direct causes and direct effects of the target of interest, and (b) discovery of the Markov blanket of the target of interest. For the first task, two major algorithms HITON_PC and MMPC were introduced by Aliferis et al. [2]. For the second task, the discovery of the Markov blanket of a target is to find the set of parents, children, and parents of the children for the target of interest in a faithful Bayesian network. Margaritis and Thrun first invented a sound algorithm, GS for discovery of the Markov blanket of a target [27]. Based on the GS algorithm, an IAMB algorithm was presented which guarantees to find the actual Markov blanket given enough training data and is more efficient than GS [1]. However, it still requires a sample size exponential in the size of the Markov blanket. Based on the IAMB algorithm, HITON_MB derived from HITON_PC, MMMB developed from MMPC, and PCMB have been introduced without requiring a sample set exponential to the size of the Markov blanket [2, 30].

3 Definitions and Notations

3.1 Emerging Patterns

Assume we have a dataset D defined upon a set of N features $(F_1, F_2, \ldots, F_N)$ and the class attribute C. For every feature F_i, $i = 1, \ldots, N$, we assume it is in a discrete domain that we denote as $\mathrm{dom}(F_i)$. Let I be the set of all items, $I = \bigcup_{i=1}^{N} \mathrm{dom}(F_i)$. An itemset X is a subset of I and its support in D, denoted $\mathrm{support}_D(X)$, is defined as follows.

Definition 1. (Support) $support_D(X) = \frac{count_D(X)}{|D|}$ (1)

where $\mathrm{count}_D(X)$ is the number of instances in D containing X and |D| is the number of instances in D.

Let $C = \{C_1, C_2, \ldots, C_K\}$ be a finite set of K distinct class labels. The dataset D can be partitioned into $D_1, D_2, \ldots, D_K$, where D_j consists of instances with class label C_j, $j = 1, \ldots, K$. The growth rate of X from D_s to D_m $(s, m = 1, \ldots, K$ and $s \neq m)$ is defined as follows.

Definition 2. (GR: Growth Rate)[9] $\mathrm{GR}_{\mathrm{D_s}\to\mathrm{D_m}}(\mathrm{X}) = \frac{\mathrm{support}_{\mathrm{D_m}}(\mathrm{X})}{\mathrm{support}_{\mathrm{D_s}}(\mathrm{X})}$. (1) If $\mathrm{support}_{\mathrm{D_m}}(\mathrm{X}) = 0$ and $\mathrm{support}_{\mathrm{D_s}}(\mathrm{X}) = 0$, then $\mathrm{GR}_{\mathrm{D_s}\to\mathrm{D_m}}(\mathrm{X}) = 0$; and (2) if $\mathrm{support}_{\mathrm{D_m}}(\mathrm{X}) \neq 0$ but $\mathrm{support}_{\mathrm{D_s}}(\mathrm{X}) = 0$, then $\mathrm{GR}_{\mathrm{D_s}\to\mathrm{D_m}}(\mathrm{X}) = \infty$.

Definition 3. (EP: Emerging Pattern)[9] Given a threshold $\rho > 1$, an EP from $\mathrm{D_s}$ to $\mathrm{D_m}$ is an itemset X where $\mathrm{GR}_{\mathrm{D_s}\to\mathrm{D_m}}(\mathrm{X}) \geq \rho$.

Definition 4. (JEP: Jumping Emerging Pattern) If $\mathrm{GR}_{\mathrm{D_s}\to\mathrm{D_m}}(\mathrm{X}) = \infty$, the itemset X is called a Jumping EPfrom $\mathrm{D_s}$ to $\mathrm{D_m}$.

An EP e from $\mathrm{D_s}$ to $\mathrm{D_m}$ is called an EP e of $\mathrm{D_m}$. The goal of EP mining is to extract the EP set $\mathrm{E_i}$ for each class $\mathrm{C_i}$ which consists of EPs from $\mathrm{D} - D_{C_i}$to $D_{C_i'}$ given a pre-defined growth rate threshold ρ and a minimum support threshold.

Definition 5. (Growth Rate Improvement)[37] Given an EP e, the growth rate improvement of e, Rateimp(e), is defined as the minimum difference between its growth rate and the growth rates of all of its subsets,

$$\mathrm{Rateimp(e)} = \min\left(\forall \mathrm{e}' \subset \mathrm{e}, \mathrm{GR(e)} - \mathrm{GR(e')}\right). \quad (2)$$

Definition 5 illustrates that a positive growth rate improvement threshold, $\mathrm{Rateimp(e)} > 0$, ensures a concise and representative set of EPs that are not subsumed by each other and consist of EPs with strong predictive power. Thus, the growth rate improvement can help to eliminate EPs that are uninteresting or redundant. Table 3 shows the Balloon dataset with the class attribute, *inflated* (T (true), F (false)), 4 features: *color* (yellow, purple), *size* (large, small), *act* (stretch, dip) and *age* (adult, child), 20 samples from the UCI machine learning repository [4], and act-r which is an artificial feature added by us that is redundant to *act*.

TABLE 3 THE BALLOON DATASET WITH AN INCLUSION OF REDUNDANT FEATURE *act-r*

ID	color	size	act-r	act	age	Inflated
1	yellow	small	yes	stretch	adult	T
2	yellow	small	yes	stretch	child	T
3	yellow	small	no	dip	adult	T
4	yellow	large	yes	stretch	adult	T
5	yellow	large	yes	stretch	child	T
6	yellow	large	no	dip	adult	T
7	purple	small	yes	stretch	adult	T
8	purple	small	yes	stretch	child	T
9	purple	small	no	dip	adult	T
10	purple	large	yes	stretch	adult	T
11	purple	large	yes	stretch	child	T
12	purple	large	no	dip	adult	T
13	yellow	small	no	dip	child	F
14	yellow	small	no	dip	child	F
15	yellow	large	no	dip	child	F
16	yellow	large	no	dip	child	F
17	purple	small	no	dip	child	F
18	purple	small	no	dip	child	F
19	purple	large	no	dip	child	F
20	purple	large	no	dip	child	F

An illustrating example is given in Tables 4 and 5 using the Balloon dataset. The minimum support threshold is 0.2 and the growth rate threshold is $\rho > 1$. The candidate EPs are of two classes T (when the inflated is true) and F (when the inflated is false) with 20 samples and 4 features: *color*, *size*, *act* and *age*.

TABLE 4 THE CANDIDATE EPS FROM CLASS F TO CLASS T

Candidate EP	Support (class F)	Support (class T)	$GR_{F\to T}$(e)
{act=stretch}	0	0.67	∞
{age=adult}	0	0.70	∞

TABLE 5 THE CANDIDATE EPS FROM CLASS T TO CLASS F

Candidate EP	Support (class T)	Support (class F)	$GR_{T\to F}$(e)
{act=dip}	0.33	1	3
{age=child}	0.33	1	3
{act=dip, age=child}	0	1	∞

From Definition 3, in Table 4, both {act=stretch} and {age=adult} are EPs of class T. In Table 5, by Definition 5, we can see that {act=dip, age=child} is an EP of class F due to $\text{Rateimp}(\text{act} = \text{dip}, \text{age} = \text{child}) > 0$.

When applying EPs to classification, the EP set of each class is used to decide to which class a test instance t should belong. More specifically, we derive k scores for t, one score per class, by feeding the EPs of each class into a scoring function, that is, $\text{label}(t) = \text{argmax}_{C_i \in C}\, \text{score}(t, C_i)$. The following definition provides the scoring function of the EPs-based classifier [10].

Definition 6 (Aggregate Score). Given an instance t and a set E_i of EPs of class $C_i \in \text{dom}(C)$ mined from the training data, the aggregate score of t for C_i is defined as

$$\text{score}(t, C_i) = \sum_{e \subseteq t, e \in E_i} \frac{\text{GR}_{D-D_{C_i}\to D_{C_i}}(e)}{\text{GR}_{D-D_{C_i}\to D_{C_i}}(e)+1} * \text{support}_{C_i}(e) \qquad (3)$$

A potential problem in Definition 6 is that the number of EPs from different classes is likely unbalanced. If a class C_i contains more EPs than another class C_j, a test instance tends to obtain higher scores for C_i than for C_j, even if the test instance actually belongs to C_j. Thus, the score computed by Definition 6 cannot be directly used to classify a test instance. Dong et al. [11] presented a concept of a base score for class C_i, $\text{baseScore}(C_i)$, which was first calculated from the training instances of the class. With the base score, the new score of an instance t for C_i, named $\text{normScore}(t, C_i)$, is defined as the ratio of the score, $\text{score}(t, C_i)$, calculated by Definition 6 and the base score, $\text{baseScore}(C_i)$,

$$\text{normScore}(t, C_i) = \frac{\text{score}(t, C_i)}{\text{baseScore}(C_i)}. \qquad (4)$$

The class with the highest normScore wins and ties are broken by putting the test instance into the class with the largest population. One way to determine the base scores is that $\text{baseScore}(C_i)$ can be the median of the scores of the training instances of class C_i [11]. For example, assume there are 5 training instances from each of the positive (+) and negative (-) classes; with all EPs of each class, assume the scores of the positive training instances computed by Definition 6 are 17.85, 18.61, 18.76, 19.75, 20.24, and the scores of the negatives are 7.8, 7.87, 8.20, 8.57, 8.61. The (median) base scores for the positive and negative classes are 18.76 and 8.20, respectively. Given a test instance t (known to be from the

negative class) with scores 10.17 and 7.92 for the positive and negative classes respectively, we have normScore(t, +)= 10.17/18.76 = 0.54 and normScore(t, −)= 7.92/8.2 =0.97. The instance s is thus labeled as the negative class.

Later, Zhang et al. introduced a simpler score function based on information theory to avoid computing the base score for each class [38] and defined the score function of a test instance t by Eq.(5).

$$L(t||C_i) = -\sum_{k=1}^{p} \log_2 P(X_k|C_i), X_k \in E_i \text{ and } X_k \in t \quad (5)$$

The test instance t is assigned the class label C_i when $L(t||C_i)$ is the minimum. Given an itemset X, $P(X|C_i)$ is approximately computed by Eq. (6).

$$P(X|C_i) = (|X \cap C_i| + 2 * (\frac{|X|}{|D|}))/(|C_i| + 2) \quad (6)$$

where $|X \cap C_i|$ is the number of training instances belonging to class C_i and containing X, $|X|$ is the total number of training instances containing X, $|D|$ is the total number of training instances, and $|C_i|$ is the number of training instances of class C_i. In addition, to ensure that we can always find a partition for an instance, all single-item itemsets of each class whether they satisfy the given thresholds or not are taken into account when Eq. (5) is used to classify a test instance.

3.2 Causal Bayesian Networks

Discovery of causal relationships between events has found wide applications in science and technology. Since late 1980's, the work on formal theories of causality and causal induction by Spirtes, Pearl and others has been gaining ground [29, 31]. Since causal Bayesian networks provide a convenient framework for reasoning among random variables, to simplify our presentation, we focus on causal Bayesian networks to represent causal relationships between variables in this paper. Since a causal Bayesian network is a Bayesian network and its structural learning in observational data is essentially the same as structure learning of Bayesian networks, one of the most exciting prospects in the last two decades has been the possibility of using Bayesian networks to discover causal relationships among features in observed data [29, 31]. The words "node" and "feature" are used interchangeably in the rest of this paper.

Definition 7 (Bayesian Networks) Let P be a discrete joint probability distribution of a set of random nodes F via a directed acyclic graph G. We call the triplet $< F, G, P >$ a (discrete) Bayesian network if $< F, G, P >$ satisfies the Markov condition: every node is independent of any subset of its non-descendant nodes conditioned on its parents.

With the Markov condition, a Bayesian network encodes the joint probability P over a set of nodes $F = \{F_1, F_2, \dots, F_n\}$ and decomposes the joint probability into a product of the conditional probability distributions over each node given its parents in G. Assuming $Pa(F_i)$ is the set of parents of F_i in G, the joint probability P is written as Eq. 7.

$$P(F_1, F_2, \dots, F_n) = \prod_{i=1}^{n} P(F_i|Pa(F_i)) \quad (7)$$

A simple Bayesian network is shown in Fig. 1 [14]. The number of possible values each node can take and the probabilities that are associated with this structure are not shown for better clarity.

Definition 8 (Faithfulness) A Bayesian network satisfies the faithfulness condition if and only if every conditional independence entailed by the directed acyclic graph G is also present in the joint probability P.

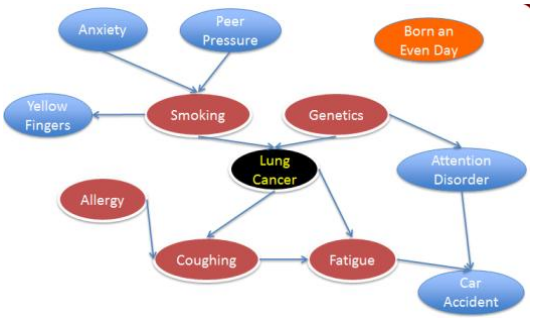

Fig. 1. A simple example of a Bayesian network of Lung Cancer

Definition 9 (Causal Bayesian Networks) A causal Bayesian network is a Bayesian network$< F, G, P >$ with the additional semantics that for all $F_i \in F$ and $F_j \in F, i \neq j$, if a node F_i is a parent of node F_j in G, then F_i is a direct cause for F_j.

Definition 10 (Causal Markov Condition) In a causal Bayesian network, if every node is independent of its non-effects (i.e., non-descendants) given its direct causes (i.e., parents), then the causal Markov condition holds.

The causal Markov condition permits the joint distribution of the features in a causal Bayesian network to be factored as in Eq. 7.

Definition 11 (Causal Faithfulness) A causal Bayesian network satisfies the faithfulness condition if it satisfies the faithfulness condition of Definition 8.

4 A FRAMEWORK OF MINING EPS FROM HIGH-DIMENSIONAL DATA

4.1 Causal Relevance and EP Discriminability

It is infeasible to examine a search space covering all possible item combinations for high-dimensional and dense data. A potentially effective way to mine EPs from high-dimensional data is to avoid the combinations of all items. From Tables 4 to 5, we can see that the final set of EPs does not contain features *size* and *color*, since their corresponding EPs have no impact on the construction of accurate classifiers. Motivated by this observation, in this section, we bridge causal relevance in causal Bayesian networks and EP discriminability (the predictive ability of an EP) in EP mining to address the two challenges on the minimal strongly predictive EP set and the impact of the minimal support threshold.

With the causal Markov condition, we define the causal relevance of a target node in causal Bayesian networks with respect to other nodes in three categories, irrelevant nodes, Markov blanket, and redundant nodes as follows.

Definition 12 (Irrelevant Nodes) In a causal Bayesian network, if F_i has no paths to connect with the target node C, node F_i is an irrelevant node with respect to C, that is,

$$\forall f \in dom(F_i), \forall c \in dom(C), P(C = c|F_i = f) = P(C = c) \quad (8)$$

In causal Bayesian networks, if a node F_i has no path to a target node C, it doesn't carry any predictive information about C at all, no matter what the context is. For example, let the node "*Lung Cancer*" be a target node in Fig. 1., node "*Born an Even Day*" in Fig. 1 is disconnected from "*Lung Cancer*", thus the pattern: {"*Born an Even Day*"=*yes*} or {"*Born an Even Day*"=*no*} cannot provide any predictive information to the target node of Lung Cancer.

Definition 13 (MB: Markov Blanket)[29] In a causal Bayesian network, the Markov blanket of a node F_i, denoted as $MB(F_i)$, is the set of its direct causes, its direct effects and the direct causes of its direct effects (spouses).

For example, in Fig.1, the Markov blanket of node "*Lung Cancer*" includes direct causes: "*Smoking*" and "*Genetics*", direct effects: "*Coughing*" and "*Fatigue*", and direct cause of the direct effects (spouse): "*Allergy*".

Property 1[29] In causal Bayesian networks with causal faithfulness, the $MB(F_i)$ is unique and satisfies the following property:

$$\forall S \in F - (MB(F_i) \cup \{F_i\}),\ P(F_i|MB(F_i), S) = P(F_i|MB(F_i)) \quad (9)$$

This property says that the Markov blanket of a node F_i is not only unique but also stores information about F_i that cannot be obtained from any other nodes in causal Bayesian networks. For example, in Fig.1, if we know the information of the Markov blanket of "*LungCancer*", it shields "*LungCancer*" from other nodes. Thus, if we know the Markov blanket of "*LungCancer*", any nodes outside of it would be redundant. The redundant nodes in causal Bayesian networks are defined as follows.

Definition 14 (Redundant Nodes) In a causal Bayesian network, if a node F_i has a path to connect with the target node C but doesn't belong to MB(C), then it is a redundant node with respect to C.

In a causal Bayesian network, if a node is redundant with respect to a target node C, the values of this node are fully determined by the MB(C). For example, with the causal Bayesian network in Fig.1, according to the causal Markov condition (see Definition 10), once all the direct causes of node "*LungCancer*" have been given, the values of its indirect causes are fully determined by their corresponding direct causes of node "*LungCancer*". Thus, the indirect causes of "*Lung Cancer*" don't bring any additional information to "*Lung Cancer*". For instance, increased "*Anxiety*" will increase "*Smoking*," but this cannot influence directly "*LungCancer*," when the value of "*Smoking*" is known in advance. Consequently, with two patterns for predicting whether a person suffers from lung cancer, *{"Smoking"=yes}→{"Lung Cancer"=yes}* and *{"Anxiety"=yes and "Smoking"=yes}→{"Lung cancer"=yes},* from Fig. 1, it suffices to have *{"Smoking"=yes}→ {"Lung Cancer"=yes}* as a predictive pattern, and we do not need

to know about "*Anxiety*." With the Markov blanket of a target node, other nodes in a causal Bayesian network become irrelevant or redundant nodes with respect to the target node.

In Fig. 2a, this Bayesian network is learned from the Balloon dataset in Table 3 without considering the artificial feature *act_r* (using the MMHC algorithm with the parameter alpha=0.01 [34]). We can see that both *color* and *size* are irrelevant to the class attribute *inflated* (in red color) while features *act* and *age* are both direct causes of the class attribute. In fact, in Tables 4 and 5, the EPs of both classes don't include features *color* and *size*. In Figure 2b, the Bayesian network is learned from the Balloon dataset with the artificial feature *act_r* that is redundant to *act*, as shown in Table 3. We can see that feature *act_r* is also a redundant node with respect to the class attribute *inflated*.

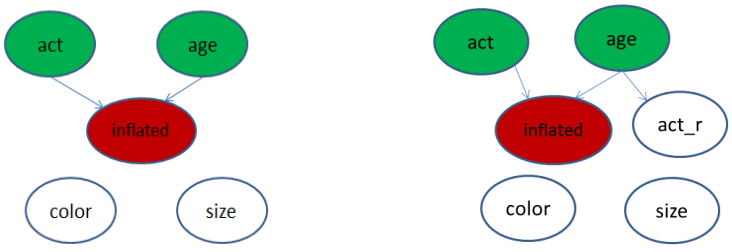

Fig. 2 (a) The Bayesian network learned from the Balloon dataset; (b) the Bayesian network learned from the Balloon dataset with the artificial feature act-r.

The above observations further motivate us to explore the potential relationships between causal relevance in a causal Bayesian network and EP discriminability in EP mining, as shown in Fig. 3, to handle EP mining from high-dimensional data. We give the following propositions to address these relationships.

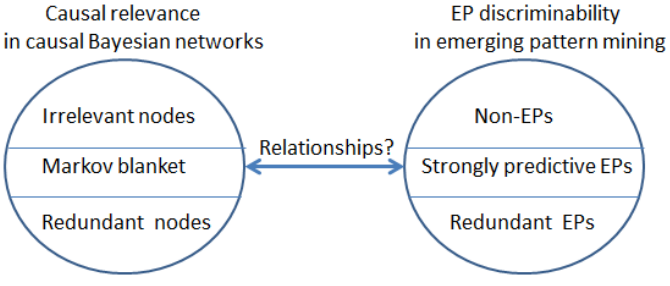

Fig.3.Causal relevance and EP discriminability

Proposition 1. If F_i is an irrelevant node with respect to the target node C in a causal Bayesian network, then $\forall f \in dom(F_i)$, the pattern $\{F_i = f\}$ is a non-EP.

Proof. Assume a dataset D has two classes $C = \{C_p, C_n\}$, D_p represents C_p class data, D_n represents C_n class data, $sup_{D_p}(F_i = f)$ is the support value of the itemset $\{F_i = f\}$ in D_p and $sup_{D_n}(F_i = f)$ is its support value in D_n. Then $GR(F_i = f)$ from D_n to D_p is calculated as follows.

$$\begin{aligned} GR_{D_n \to D_p}(F_i = f) &= \sup_{D_p}(F_i = f) / \sup_{D_n}(F_i = f) \\ &= \mathrm{P}(F_i = f | \mathrm{C} = \mathrm{C_p}) / \mathrm{P}(F_i = f | \mathrm{C} = \mathrm{C_n}) \\ &= \frac{\mathrm{P}(F_i = f, \mathrm{C} = \mathrm{C_p})}{\mathrm{P}(\mathrm{C} = \mathrm{C_p})} \Big/ \frac{\mathrm{P}(F_i = f, \mathrm{C} = \mathrm{C_n})}{\mathrm{P}(\mathrm{C} = \mathrm{C_n})} \\ &= \frac{\mathrm{P}(\mathrm{C} = \mathrm{C_p} | F_i = f)\mathrm{P}(F_i = f)}{\mathrm{P}(\mathrm{C} = \mathrm{C_p})} \Big/ \frac{\mathrm{P}(\mathrm{C} = \mathrm{C_n} | F_i = f)\mathrm{P}(F_i = f)}{\mathrm{P}(\mathrm{C} = \mathrm{C_n})} \\ &= \frac{\mathrm{P}(\mathrm{C} = \mathrm{C_n})}{\mathrm{P}(\mathrm{C} = \mathrm{C_p})} \cdot \frac{\mathrm{P}(\mathrm{C} = \mathrm{C_p} | F_i = f)}{\mathrm{P}(\mathrm{C} = \mathrm{C_n} | F_i = f)} \end{aligned}$$

Since $\mathrm{F_i}$ is an irrelevant node with respect to the target node C in the causal Bayesian network, by Eq.(8), we get the following equation.

$$\begin{aligned} GR_{D_n \to D_p}(F_i = f) &= \frac{1\text{-}\mathrm{P}(\mathrm{C} = \mathrm{C}_p)}{\mathrm{P}(\mathrm{C} = \mathrm{C}_p)} \cdot \frac{\mathrm{P}(\mathrm{C} = \mathrm{C_p} | F_i = f)}{1\text{-}\mathrm{P}(\mathrm{C} = \mathrm{C_p} | F_i = f)} \\ &= 1 \end{aligned}$$

According to Definition 3 in Section 2.1, Proposition 1 is proven. □

From Definition 5 in Section 2.1, for an EP e, if we can find an $\mathrm{e}' \subset \mathrm{e}$ to make $\mathrm{Rateimp(e)} \leq 0$, then e might be an uninteresting or redundant EP given its subset e', and the EP e might be replaced by its subset e'. Thus, avoiding generation of these redundant EPs in advance will improve search efficiency. We give Proposition 2 below to explain the relationships between redundant nodes in causal Bayesian networks and EP redundancy in EP mining.

Proposition 2. If a node $\mathrm{F_i}$ is a redundant node to the target node C in a causal Bayesian network and M is the Markov blanket of C, then $\forall \mathrm{f} \in \mathrm{dom(F_i)}$, and there exists $\mathrm{m} \in \bigcup_{\mathrm{i}=1}^{|\mathrm{M}|} \mathrm{dom(M_i)}$, such that the candidate EP of $\{\mathrm{F_i} = \mathrm{f}, \mathrm{M} = \mathrm{m}\}$ is a redundant EP with respect to the EP of $\{\mathrm{M} = \mathrm{m}\}$.

Proof. Since $\mathrm{F_i}$ is a redundant node with respect to C, by Property 1, the following equation holds:

$$\forall \mathrm{c} \in \mathrm{dom(C)}, \mathrm{P(C = c | M = m, F_i = f) = P(C = c | M = m)} \qquad (10)$$

$GR(\mathrm{F_i = f, S = s})$ from $\mathrm{D_n}$ to $\mathrm{D_p}$ is calculated as follows.

$$\begin{aligned} GR_{D_n \to D_p}\left(F_i = \mathrm{f}, M = m\right) &= \sup_{D_p}(F_i = f, M = m) / \sup_{D_n}(F_i = f, M = m) \\ &= P(F_i = \mathrm{f}, M = m | \mathrm{C}{=}\mathrm{C}_p) / P(F_i = \mathrm{f}, M = m | \mathrm{C}{=}\mathrm{C}_n) \\ &= \frac{P(\mathrm{C}{=}\mathrm{C}_p | F_i = \mathrm{f}, M = m)}{P(\mathrm{C}{=}\mathrm{C}_p)} \Big/ \frac{P(\mathrm{C}{=}\mathrm{C}_n | F_i = \mathrm{f}, M = m)}{P(\mathrm{C}{=}\mathrm{C}_n)} \\ &= \frac{P(\mathrm{C}{=}\mathrm{C}_p | F_i = \mathrm{f}, M = m)}{P(\mathrm{C}{=}\mathrm{C}_n | F_i = \mathrm{f}, M = m)} \cdot \frac{P(\mathrm{C}{=}\mathrm{C}_n)}{P(\mathrm{C}{=}\mathrm{C}_p)} \end{aligned}$$

Since $\mathrm{F_i}$ is a redundant node to C in the causal Bayesian network, according to Eq.(10), we get the following equation.

$$\begin{aligned} GR_{D_n \to D_p}\left(F_i = \mathrm{f}, M = m\right) &= \frac{1 - P(\mathrm{C}{=}\mathrm{C}_n | F_i = \mathrm{f}, M = m)}{P(\mathrm{C}{=}\mathrm{C}_n | F_i = \mathrm{f}, M = m)} \cdot \frac{P(\mathrm{C}{=}\mathrm{C}_n)}{P(\mathrm{C}{=}\mathrm{C}_p)} \\ &= \frac{1 - P(\mathrm{C}{=}\mathrm{C}_n | M = m)}{P(\mathrm{C}{=}\mathrm{C}_n | M = m)} \cdot \frac{P(\mathrm{C}{=}\mathrm{C}_n)}{P(\mathrm{C}{=}\mathrm{C}_p)} \\ &= \frac{P(\mathrm{C}{=}\mathrm{C}_p | M = m)}{P(\mathrm{C}{=}\mathrm{C}_n | M = m)} \cdot \frac{P(\mathrm{C}{=}\mathrm{C}_n)}{P(\mathrm{C}{=}\mathrm{C}_p)} \\ &= \frac{P(M = m | \mathrm{C}{=}\mathrm{C}_p)}{P(M = m | \mathrm{C}{=}\mathrm{C}_n)} \\ &= GR_{D_n \to D_p}(M = m) \end{aligned}$$

Thus, we get $GR_{D_{n\to D_p}}(\mathrm{F_i}=\mathrm{f},\mathrm{M}=\mathrm{m})-GR_{D_{n\to D_p}}(\mathrm{M}=\mathrm{m})=0$. According Definition 5 in Section 2.1, we have proven Proposition 2. □

With the results of Propositions 1 and 2, we can remove irrelevant and redundant nodes with respect to the class attribute in causal Bayesian networks to achieve the goal of pruning non-EPs and redundant EPs before EP mining. Thus, by removing irrelevant and redundant nodes in causal Bayesian networks, we can reduce the pattern space in EP mining to the space of the Markov blanket of the class attribute in a causal Bayesian network, and we get Proposition 3 as follows.

Proposition 3. The pattern space in EP mining for classification can be reduced to the space of the Markov blanket of the class attribute in causal Bayesian networks.

With Proposition 3, within the Markov blanket of the class attribute, both the direct causes and the direct effects have a direct connecting path to the class attribute while the direct causes of the direct effects (spouses) of the class attribute don't. Thus, in causal Bayesian networks, the spouses of the class attribute cannot individually predict the class attribute and may enhance the predictive power of the direct effects only when they join with the direct effects. With the causal Markov condition, direct causes joined with the direct effects give the highest predictive ability to predict the class attribute and a natural interpretation for what happens to the class attribute. For example, as indicated in Fig.1, the nodes "*Smoking*", "*Genetics*", "*Coughing*", and "*Fatigue*" give the most highly predictive ability to predict whether a person suffers from "*Lung Cancer*", while the node "*Allergy*" cannot individually predict "*Lung Cancer*". "*Allergy*" may enhance the predictive power of "*Coughing*" when it joins with "*Coughing*". Thus, the pattern space in EP mining can be naturally reduced to direct causes and direct effects of the class attribute in causal Bayesian networks, and then we obtain Proposition 4 as follows.

Proposition 4. The pattern space in EP mining for classification can be further reduced to direct causes and direct effects of the class attribute in causal Bayesian networks.

4.2 Mining Emerging Patterns from High-dimensional Data

With Propositions 1 to 4 above, we give a new framework for mining EPs with high-dimensional data, as shown in Fig. 4 with 4 steps. The key steps of our framework are (1) Step 1: identifying CE (direct ***C***auses and direct ***E***ffects) or MB (the ***M***arkov ***B***lanket) of the class attribute from data, and (2) Step 3: mining EPs from CE or MB space.

Step 1: Identifying CE and MB of the class attribute. As stated in Section 2.2, structure learning of causal Bayesian networks in observational data is essentially the same as structure learning of Bayesian networks. When the number of features is small, we can adopt existing Bayesian network structure learning algorithms to construct a complete Bayesian network, and then get the CE or MB of the class attribute. When there are tens of thousands of feature dimensions, learning a complete Bayesian network is simply impossible [8].

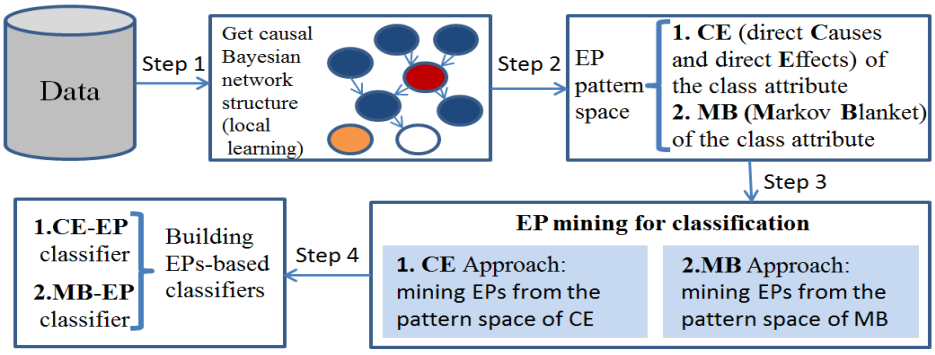

Fig. 4 The framework of building EPs-based classifiers from high-dimensional data

In fact, in our framework, both the CE approach and the MB approach only need to get the CE or MB of the class attribute, and don't need to distinguish which node is a direct cause, a direct effect, or a spouse from the CE set or MB set. Rather than recovering a complete Bayesian network among all features, our framework only uses local learning techniques to uncover cause-effect relationships between the class attribute and other nodes or to uncover the MB of the class attribute. Most importantly, in a causal Bayesian network with causal faithfulness, the CE or MB of a target is unique and minimal.

Therefore, no matter whether the number of features in a dataset is small or large, our framework adopts local learning techniques to capture the CE or the MB of the class attribute. For the CE approach of EP mining, there are two state-of-the-art local learning techniques, MMPC and HITON_PC (detailed descriptions in [2]). Since those two algorithms are complete under the assumption of causal faithfulness, we introduce the HITON_PC algorithm into our framework to get the CE of the class attribute without uncovering a complete Bayesian network. For the MB approach, the HITON_MB algorithm is used to get the MB of the class attribute [2][1]. To identify the MB of the class attribute, the HITON_MB algorithm first discovers the CE of the class attribute by the HITON_PC algorithm and, then, identifies the spouses of the class attribute.

Step 3: Mining EPs from the pattern space of CE or MB of the class attribute. At step 3, with the CE or MB of the class attribute, our framework gives two approaches, CE approach to mine EPs from the space of direct causes and direct effects, and MB approach to mine EPs from the space of the Markov blanket. Since the CE or MB of the class attribute is unique and minimal in a causal Bayesian network, at step 3, we adopt the ConsEPMiner algorithm[2] which is a level-wise, candidate generation-and-test approach to mine EPs [37]. The ConsEPMiner algorithm follows the set-enumeration tree search framework and the breadth-first search strategy, and mines EPs satisfying several constraints including the growth-rate improvement constraint. With the EPs mined by our two approaches, two classifiers, the CE-EP and MB-EP classifiers, are constructed, and they both use the score function defined by Eq.5 in Section 3.1 for classification.

[1] The codes of HITON_PC and HITON_MB are available at http://www.dsl-lab.org/causal_explorer.
[2] The code of the ConsEPMiner algorithm is available at http://goanna.cs.rmit.edu.au/~zhang.

5 EXPERIMENTAL RESULTS

5.1 Experimental Setup

In order to thoroughly evaluate the proposed framework, thirty six datasets (Table 6) are selected, including the UCI datasets (the first 24 datasets), very high-dimensional biomedical datasets (*hiva*, *ovarian-cancer*, *lymphoma*, and *breast-cancer*), NIPS 2003 feature selection challenge datasets (*madelon*, *arcene*, *dorothea*, and *dexter*), and four frequently studied public microarray datasets (the last 4 datasets), respectively.

TABLE 6 #: THE NUMBER OF FEATURES, SIZE: THE NUMBER OF INSTANCES

Dataset	#	SIZE	Dataset	#	SIZE
australian	14	690	promoters	57	106
breast-w	9	3,146	spect	22	267
crx	15	690	spectf	44	267
cleve	13	303	tictactoe	9	958
diabetes	8	768	vote	16	435
german	20	1,000	wdbc	30	569
house-votes	16	230	madelon	500	2,000
hepatitis	19	155	hiva	1,617	4,229
horse-colic	22	368	ovarian-cancer	2,190	216
hypothyroid	25	3,163	lymphoma	7,399	227
heart	13	270	dexter	20,000	300
infant	86	5,337	breast-cancer	17,816	286
ionosphere	34	351	arcene	10,000	100
kr-vs-kp	36	3,196	dorothea	100,000	800
labor	16	57	colon	2,000	62
liver	6	345	leukemia	7,129	72
mushroom	22	8,124	lung-cancer	12,533	181
pima	8	768	prostate	6,033	102

Our comparative study involves three types of comparisons, using ten-fold cross-validation on all datasets.

- Comparing CE-EP and MB-EP classifiers against a well-known EP classifier, CAEP [10] and a Strong Jumping EP classifier, SJEP [11].
- Comparing CE-EP and MB-EP classifiers with three well-known associative classifiers: CBA [22], CMAR [19] and CPAR [36].
- Comparing CE-EP and MB-EP classifiers with the state-of-the-art non-associative classifiers, including Naïve Bayes (NB), Knn, Decision Tree J48, SVM, Bagging and AdaBoost using their Weka implementations with default parameters [15].

To discretize continuous features, we use the discretization method in the Causal Explorer Toolkit provided by Aliferis et al. [1]. In the experiments, we set the minimum confidence threshold to 0.8 for CBA and CMAR, and set the growth rate to 20 for CAEP, CE-EP and MB-EP classifiers. To thoroughly test the impact of the support threshold values, we set seven minimum supports for CE-EP, MB-EP, CAEP, CBA, and CMAR, including 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, and 0.4, respectively. We select the best classification accuracy under the seven minimum supports as the results for our comparative study. The SJEP classifier uses the minimal support threshold suggested in the original paper [11]. The parameters

for CPAR are set the same as those reported in [36]. CBA, CMAR, and CPAR are implemented in the LUCS KDD Software Library in Java [26], while CE-EP, MB-EP and CAEP are implemented in C++. The experiments are performed on a Window 7 DELL workstation with an Intel Xeon 2.93 GHz processor and 12.0 GB RAM.

5.2 Comparison of Predictive Accuracy

5.2.1. Comparing with Other 11 Classifiers

Tables 7 to 9 report the predictive accuracies of our two classifiers, CE-EP and MB-EP, in comparison with other eleven classifiers, including two EP, three associative and six non-associative classifiers on the thirty six benchmark datasets. We select the best predictive accuracy under the seven minimum supports as the results for our comparative study. The best results among all classifiers are highlighted in bold for each dataset and the symbol "/" denotes that the classifier runs out of memory due to a huge number of candidate patterns.

In Table 7, compared to CAEP and SJEP, CE-EP achieves the highest accuracy on twenty one datasets out of the thirty six datasets while CAEP and SJEP fail to deal with the twelve very high-dimensional datasets. On three datasets, *diabetes*, *liver* and *pima*, SJEP gets very low predictive accuracy, since there are not enough SJEPs in those datasets for classification. In Table 8, CBA, CMAR and CPAR only have results on the twenty four low dimensional datasets as they fail to deal with a high feature space. From Tables 7 and 8, we can see that CE-EP outperforms SJEP, CBA, CMAR, and CPAR. Table 9 compares the accuracy of our two classifiers against well-known classifiers such as Naive Bayes (NB), Decision Tree J48, Knn, SVM and two ensemble classifiers, Bagging and AdaBoost. In comparison with these six classifiers, CE-EP is significantly superior to NB, Knn, J48, Bagging and AdaBoost and very competitive with SVM on all the thirty six datasets in Table 9.

TABLE 7 COMPARISON OF PREDICTIVE ACCURACY (%): CE-EP, MB-EP, SJEP, AND CAEP

Dataset	CE-EP	MB-EP	SJEP	CAEP	Dataset	CE-EP	MB-EP	SJEP	CAEP
australian	83.97	83.97	78.24	**84.71**	promoters	**72.00**	**72.00**	/	/
breast-w	**96.88**	**96.88**	90.80	**96.88**	spect	**72.69**	**72.69**	56.92	69.23
crx	82.21	82.21	76.32	**84.85**	spectf	83.85	**85.00**	75.00	/
cleve	84.83	82.76	81.72	**85.52**	tictactoe	69.58	69.58	**99.88**	82.95
diabetes	**72.11**	71.18	20.79	68.95	vote	**95.95**	**95.95**	93.33	90.00
german	71.50	71.30	**73.90**	72.80	wdbc	81.79	**83.39**	70.36	81.96
house-votes	**96.82**	**96.82**	93.18	90.91	madelon	59.00	**60.85**	/	/
hepatitis	85.33	85.33	**86.00**	**86.00**	hiva	**93.70**	93.67	/	/
horse-colic	**85.83**	83.33	81.39	79.72	ovarian-cancer	**92.86**	**92.86**	/	/
hypothyroid	72.82	72.82	**75.47**	67.78	lymphoma	**77.73**	**77.73**	/	/
heart	83.70	82.96	81.85	**85.93**	dexter	88.33	**89.33**	/	/
infant	**94.92**	94.78	/	/	arcene	**86.67**	**86.67**	/	/
ionosphere	92.94	91.76	**94.74**	90.29	breast-cancer	**92.22**	91.48	/	/
kr-vs-kp	**92.23**	91.54	91.82	83.49	dorothea	**95.06**	**95.06**	/	/
labor	**96.00**	92.00	86.00	94.00	colon	**95.00**	90.00	/	/
liver	**61.76**	**61.76**	10.29	57.65	leukemia	**100.00**	**100.00**	/	/
mushroom	96.18	95.54	**98.12**	96.18	lung-cancer	99.44	**100.00**	/	/
pima	**72.11**	71.18	20.79	68.95	prostate	**94.00**	**94.00**	/	/

TABLE 8 COMPARISON OF PREDICTIVE ACCURACY (%): CE-EP, CBA, CMAR, AND CPAR

Dataset	CE-EP	CBA	CMAR	CPAR	Dataset	CE-EP	CBA	CMAR	CPAR
australian	83.97	**86.96**	**86.96**	85.51	ionosphere	**92.94**	88.88	90.58	88.88
breast-w	**96.88**	94.09	90.82	92.95	kr-vs-kp	92.23	**93.56**	89.41	88.71
crx	82.21	**86.52**	85.51	85.51	labor	**96.00**	54.33	89.17	80.33
cleve	84.83	83.12	**85.82**	78.61	liver	**61.76**	60.90	4.12	58.14
diabetes	72.11	73.18	64.24	**73.31**	mushroom	96.18	78.67	**99.37**	98.66
german	71.50	**74.50**	71.00	65.70	pima	72.11	**73.45**	63.94	67.97
house-votes	96.82	**96.96**	**96.96**	**96.96**	promoters	**72.00**	28.13	42.50	63.00
hepatitis	**85.33**	49.50	83.33	72.34	spect	**72.69**	64.42	62.66	64.42
horse-colic	**85.83**	83.69	83.91	82.02	spectf	**83.85**	55.84	80.07	54.74
hypothyroid	72.82	**94.78**	90.00	89.56	tictactoe	69.58	**100**	99.26	71.43
heart	83.70	**84.07**	**84.07**	77.41	vote	**95.95**	95.40	95.40	94.01
infant	**94.92**	63.72	90.00	84.30	wdbc	81.79	**95.79**	95.61	92.91

TABLE 9 COMPARISON OF PREDICTIVE ACCURACY (%) WITH NON-ASSOCIATIVE CLASSIFIERS

Dataset	CE-EP	NB	Knn	J48	SVM	Bagging	AdaBoost
australian	83.97	84.78	81.88	85.79	85.36	85.80	**86.38**
breast-w	96.88	95.99	95.14	94.56	**97.00**	95.57	94.85
crx	82.21	85.36	81.59	84.20	**85.51**	84.64	85.36
clever	**84.83**	84.82	79.54	75.91	82.51	79.54	84.82
diabetes	72.11	70.44	68.23	72.00	**73.18**	72.40	**73.18**
german	71.50	74.90	69.00	74.00	**76.20**	73.70	71.60
house-votes	**96.82**	85.19	79.63	77.03	83.33	82.96	81.11
hepatitis	**85.33**	83.87	80.65	80.65	84.52	83.23	80.65
horse-colic	**85.83**	79.08	76.34	81.79	81.79	85.05	83.70
hypothyroid	72.82	95.57	95.29	**95.64**	95.57	95.54	95.23
heart	83.70	90.87	91.74	96.52	**96.96**	**96.96**	**96.96**
infant	94.92	91.91	92.51	95.39	95.41	**95.65**	95.43
ionosphere	**92.94**	88.89	88.60	92.02	91.74	91.45	89.46
kr-vs-kp	92.23	83.92	95.15	**99.31**	94.99	99.22	93.84
labor	**96.00**	91.22	87.72	92.98	85.96	80.70	87.72
liver	**61.76**	61.16	61.74	60.00	60.29	59.71	60.87
mushroom	96.18	85.68	**100.00**	**100.00**	99.11	**100.00**	98.44
pima	72.11	70.44	68.75	72.01	**73.18**	72.40	**73.18**
promoters	72.00	74.53	55.66	63.21	**79.25**	60.38	66.04
spect	**72.69**	66.67	64.79	65.54	70.04	67.79	69.66
spectf	83.85	86.63	57.75	62.57	81.25	**90.37**	75.40
tictactoe	69.58	69.20	**98.43**	85.70	98.33	90.40	72.31
vote	**95.95**	94.23	94.23	94.01	94.93	95.40	82.87
wdbc	**81.79**	77.68	79.96	75.92	79.61	80.32	76.27
madelon	59.00	59.20	52.95	57.50	56.35	**62.20**	60.50
hiva	93.70	87.06	95.38	96.39	94.70	**96.64**	96.47
lymphoma	77.33	68.28	63.88	70.93	**77.53**	64.76	60.79
breast-cancer	92.22	**93.01**	86.36	80.77	92.31	84.97	83.57
ovarian-cancer	92.86	70.83	83.33	89.35	**93.52**	88.89	90.74
dorothea	**95.06**	90.25	90.63	89.38	92.00	94.13	93.75
arcene	**86.67**	63.00	77.00	56.00	81.00	72.00	71.00
dexter	88.33	**93.33**	63.67	82.67	91.33	86.67	81.33
colon	**95.00**	79.03	79.03	79.03	85.48	85.48	85.48
leukemia	**100.00**	93.06	97.22	90.28	98.61	90.28	**100.00**
lung-cancer	99.44	98.34	98.34	90.61	**100.00**	93.92	96.69
prostate	94.00	69.61	88.24	88.24	**94.12**	92.16	91.18

TABLE 10 WIN/TIE/LOSS COUNTS OF CE-EP VS. THE OTHER 12 CLASSIFIERS (PAIRWISE T-TEST AT 95% SIGNIFICANCE LEVEL)

	MB-EP	SJEP	CAEP	CBA	CMAR	CPAR	NB	Knn	J48	SVM	Bagging	AdaBoost
CE-EP	5/28/3	29/2/5	26/6/4	11/4/9	13/5/6	16/1/7	22/7/7	28/2/6	23/4/9	15/9/12	21/5/10	20/5/11

To further investigate the classification results, we conduct paired t-tests at a 95% significance level and summarize the win/tie/lose counts of CE-EP against the other algorithms in Table 10 (note: if a classifier fails to run on a dataset while our method can do it, then our classifier wins). With the summary of the win/tie/lose counts shown in Table 10, we can see that CE-EP usually outperforms CAEP, CBA, CMAR and CPAR. Meanwhile, the MB-EP classifier has extremely similar performance with CE-EP. In comparison with the well-known non-associative classifiers, CE-EP is significantly superior to NB, Knn,

J48, Bagging and AdaBoost, and is competitive with SVM on all the thirty six datasets.

The above empirical results demonstrate that mining EPs in the pattern space of direct causes and direct effects, or the Markov blanket of the class attribute can find high quality patterns which possess the most differentiating power. Most importantly, the CE-EP and MB-EP classifiers can not only handle very high-dimensional datasets such as the last twelve datasets in Table 6, but also produce very promising predictive accuracy.

Why do the EPs mined by CE-EP possess such high discriminating power? We use vote, a UCI dataset, as an illustrating example. The vote dataset has 16 features (attributes) and one class attribute. Each feature has two values, *yea* and *nay*, and the class attribute is divided into two classes: *Democrat* (D) and *Republican* (R). Tables 11 to 13 give the EPs of the two classes mined from the vote dataset by the CE-EP algorithm, respectively.

It is clear that the EPs in Tables 11 to 12 are constructed from features *physician-fee-freeze*, *adoption-of-the-budget-resolution* and s*ynfuels-corporation-cutback*, whose indices in the original vote dataset are 4, 3 and 11, respectively. The EPs constructed from these three features are highly discriminative as indicated by the high growth ratio GR(e). Note that Feature 4 is the direct cause of the class attribute while Features 3 and 11 are the direct effects of the class attribute in the causal Bayesian network learned from the vote dataset.

From the viewpoint of feature discriminability, the mutual information measure and the chi-squared test both show that Features 4, 3 and 11 are the most informative features among all of features with respect to the class attribute. This is consistent with the feature discriminability described in Table 13.

TABLE 11 THE EPS FROM CLASS R(REPUBLICAN) TO CLASS D (DEMOCRAT)

Emerging Patterns	Support(class R)	Support(class D)	GR(e)
{physician-fee-freeze=nay& adoption-of-the-budget-resolution=yea}	0.0179	0.8614	48.12
{physician-fee-freeze=nay&synfuels-corporation-cutback=yea}	0	0.4419	∞
{adoption-of-the-budget-resolution=yea&synfuels-corporation-cutback=yea}	0.0179	0.4082	22.81

TABLE 12 THE EPS FROM CLASS D (DEMOCRAT) TO CLASS R (REPUBLICAN)

Emerging Patterns	Support(class D)	Support(class R)	GR(e)
{physician-fee-freeze=yea&adoption-of-the-budget-resolution=nay}	0.02	0.8333	37.08
{physician-fee-freeze=yea&synfuels-corporation-cutback=nay}	0.01	0.85	75.23
{adoption-of-the-budget-resolution=nay&synfuels-corporation-cutback=nay}	0.03	0.74	21.90
{physician-fee-freeze=yea&adoption-of-the-budget-resolution=nay& synfuels-corporation-cutback=nay}	0.0037	0.73	197.3

TABLE 13 FEATURE DISCRIMINABILITY OF FEATURES 4, 3 AND 11

Feature	Feature Discriminability (Pr(C | F=V)(F denotes a feature and V means its value)	
4(physician-fee-freeze)	0.99(C=democrat; V=nay)	0.92 (C=republican; V=yea)
3 (adoption-of-the-budget-resolution)	0.91(C=democrat; V=yea)	0.83(C=republican; V=nay)
11(synfuels-corporation-cutback)	0.86(C=democrat; V=yea)	0.52(C=republican; V=nay)

Accordingly, we conclude that the EPs extracted from the pattern space of direct causes and direct effects are high-quality patterns and possess the most discriminative power. They are the best candidates to be used to construct a highly accurate classifier.

5.2.2 Comparing with Top-k Feature Ranking Methods

For the last 12 high-dimensional datasets in Table 6, we further compare the CE-EP algorithm with the two top-k feature ranking methods, the mutual information (MI for short) measure and the chi-squared test (CHI for short). For both ranking methods, we select the top 20 and top 30 features respectively, and then use the selected features to train EP classifiers to get the benchmark results.

TABLE 14 THE PREDICTIVE ACCURACY (%) OF CE-EP VS. CHI AND MI

Dataset	CE-EP	CHI(20)	CHI(30)	MI(20)	MI(30)
madelon	59.00	62.35	62.35	61.90	**63.45**
hiva	93.70	94.29	**94.43**	92.44	90.59
ovarian-cancer	**92.38**	86.77	85.71	84.29	84.76
lymphoma	77.73	74.09	77.73	**81.36**	79.55
dexter	88.33	90.67	**92.33**	85.00	88.67
breast-cancer	**92.22**	90.74	90.37	90.74	91.11
arcene	**86.67**	70.00	71.11	72.22	70.00
dorothea	**95.06**	93.92	93.92	94.05	94.18
colon	**95.00**	90.00	88.33	91.67	90.00
leukemia	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
lung-cancer	99.44	**100.00**	98.89	97.22	98.33
prostate	**94.00**	92.00	**94.00**	**94.00**	93.00
Ave.	**89.46**	87.07	87.43	87.07	86.97

For CE-EP, we set the support threshold to 0.2 and the growth rate threshold to 20. From Tables 14 to 16, we can see that CE-EP gets much better performance than both the CHI and MI methods on the predictive accuracy and the number of mined patterns (MI(20) or CHI(20) denotes the top 20 features selected by the MI or CHI method).

TABLE 15 WIN/TIE/LOSS (PAIRWISE T-TEST AT 95% SIGNIFICANCE LEVEL)

	CHI(20)	CHI(30)	MI(20)	MI(30)
CE-EP	7/3/2	5/5/2	8/2/2	8/2/2

TABLE 16 THE NUMBER OF MINED PATTERNS OF CE-EP VS. CHI AND MI

Dataset	CE-EP	CHI(20)	CHI(30)	MI(20)	MI(30)		Dataset	CE-EP	CHI(20)	CHI(30)	MI(20)	MI(30)
madelon	**22**	86	140	90	144		arcene	**27**	98	150	168	747
hiva	**28**	80	120	80	120		dorothea	**45**	107	160	507	335
ovarian-cancer	**40**	634	3480	672	2303		colon	**27**	259	815	251	807
lymphoma	**33**	827	20963	225	3111		leukemia	**54**	314	687	281	673
dexter	**35**	250	967	225	751		lung-cancer	**39**	971	3656	2636	10300
breast-cancer	**40**	234	615	1353	16287		prostate	**134**	366	1149	334	1149

5.3 Comparison of the Number of Patterns

In this section, we compare the numbers of patterns selected by CE-EP and MB-EP with the CAEP, CBA and CMAR classifiers, as these five associative classifiers all focus on generating patterns with the support-confidence framework. We report the average numbers of patterns over all seven minimum support thresholds. Since on *wdbc*, *kr-vs-kp*, *ionosphere*, *horse-colic* and *german*, CAEP cannot run using all the support thresholds due to huge numbers of patterns, the numbers of patterns on these datasets is averaged over the available support thresholds.

As depicted in Tables 17 and 18, it is clear that the CE-EP and MB-EP classifiers select many fewer patterns than the CAEP, CBA and CMAR classifiers on all the datasets.

These results further illustrate that extracting the EPs from the space of direct causes and direct effects

or the Markov blanket of the class attribute not only gets a much smaller set of EPs but also achieves a higher predictive accuracy than the existing EPs-based and associative classifiers. Tables 17 to 18 also indicate that even with very high-dimensional datasets, the numbers of patterns selected by CE-EP and MB-EP don't change much in comparison with those on the twenty four low-dimensional datasets while CAEP, CBA, and CMAR cannot deal with those datasets, even with a rather high support threshold.

TABLE 17 COMPARISON OF NUMBERS OF PATTERNS (AVERAGE ON SEVEN SUPPORT THRESHOLDS): CE-EP, MB-EP, AND CAEP

Data	CE-EP	MB-EP	CAEP	Data	CE-EP	MB-EP	CAEP
australian	**41**	**41**	751	promoters	**8**	**8**	/
breast-w	**269**	**269**	269	spect	**12**	**12**	2906
crx	**20**	**20**	866	spectf	**25**	102	/
cleve	**45**	55	449	tictactoe	**39**	**39**	227
diabetes	**20**	25	50	vote	**18**	**18**	1715
german	**30**	35	2935	wdbc	**48**	92	2840
house-votes	**10**	**10**	1041	madelon	**22**	29	/
hepatitis	**19**	24	757	hiva	**29**	114	/
horse-colic	**31**	34	2935	ovarian-cancer	**17**	**17**	/
hypothyroid	**24**	**24**	146	lymphoma	**44**	**44**	/
heart	**36**	57	331	dexter	**38**	59	/
infant-mortality	**71**	114	/	arcene	**27**	**27**	/
ionosphere	**35**	48	4160	breast-cancer	**42**	177	/
kr-vs-kp	**53**	359	1807	dorothea	**47**	65	/
labor	**37**	52	438	colon	**26**	56	/
liver	**4**	**4**	25	leukemia	**54**	2139	/
mushroom	**258**	915	1059	lung-cancer	**39**	53	/
pima	**20**	25	51	prostate	**113**	**113**	/

TABLE 18 COMPARISON OF NUMBERS OF PATTERNS (AVERAGE ON SEVEN SUPPORT THRESHOLDS): CE-EP, MB-EP, CBA, AND CMAR

Data	CE-EP	MB-EP	CBA	CMAR	Data	CE-EP	MB-EP	CBA	CMAR
australian	**41**	**41**	2462	7388	ionosphere	**35**	48	40955	18274
breast-w	**269**	**269**	477	1278	kr-vs-kp	**53**	359	6513	2746
crx	**20**	**20**	3140	5531	labor	**37**	52	8701	10650
cleve	**45**	55	3354	5570	liver	**4**	**4**	16	13
diabetes	**20**	25	205	375	mushroom	**258**	915	47214	23108
german	**30**	35	15404	1960	pima	**20**	25	208	382
house-votes	**10**	10	8356	15734	promoters	**8**	**8**	8339	272
hepatitis	**19**	24	14863	17523	spect	**12**	**12**	113	1093
horse-colic	**31**	34	25820	5304	spectf	**25**	102	11011	3489
hypothyroid	**24**	**24**	15275	5503	tictactoe	**39**	**39**	832	1005
heart	**36**	57	3068	5585	vote	**18**	**18**	8477	15734
Infant-mortality	**71**	114	109652	2972	wdbc	**48**	92	43271	27708

5.4 Comparison of Running Time of EP Classifiers

Table 19 reports the average running time over seven minimum support thresholds of CE-EP and MB-EP against CAEP. The running time contains all execution time, including importing datasets, ten-fold cross validation learning and testing. The best result for each dataset is highlighted in bold. As stated in Section 5.3, since on *wdbc*, *kr-vs-kp*, *ionosphere*, *horse-colic* and *german*, CAEP cannot run under all the support thresholds, the running time for these datasets is averaged over the available support thresholds.

Table 19 shows that CE-EP and MB-EP are faster than CAEP on all the datasets. The running time of CAEP fluctuates a little among different datasets while CE-EP and MB-EP, especially CE-EP, have a very stable running time for both low and high dimensional datasets. On the *dorothea* dataset, the running time of MB-EP is greater than CE-EP due to the very large space of this dataset, up to 100,000 features.

TABLE 19 COMPARISON OF RUNNING TIME (AVERAGE ON SEVEN SUPPORT THRESHOLDS): CE-EP, MB-EP, AND CAEP

Data	CE-EP	MB-EP	CAEP	Data	CE-EP	MB-EP	CAEP
australian	**43**	**43**	50	promoters	**30**	**30**	/
breast-w	**51**	**51**	**51**	spect	**30**	**30**	87
crx	**31**	**31**	42	spectf	**31**	32	/
cleve	**31**	**31**	53	tictactoe	**31**	**31**	33
diabetes	**31**	**31**	48	vote	**30**	**30**	47
german	**31**	32	129	wdbc	**31**	32	89
house-votes	**27**	32	54	madelon	**32**	33	/
hepatitis	**37**	39	43	hiva	**36**	37	/
horse-colic	**31**	**31**	51	ovarian-cancer	**34**	44	/
hypothyroid	**32**	**32**	107	lymphoma	**32**	**32**	/
heart	**45**	46	50	dexter	**38**	75	/
Infant-mortality	**50**	75	/	arcene	**34**	36	/
ionosphere	**43**	45	146	breast-cancer	**47**	145	/
kr-vs-kp	**48**	84	390	dorothea	**164**	1780	/
labor	**31**	**31**	42	colon	**32**	37	/
liver	**10**	**10**	**10**	leukemia	**50**	90	/
mushroom	**64**	85	100	lung-cancer	**42**	71	/
pima	**45**	**45**	46	prostate	**34**	51	/

5.5 Sensitivity Analysis on Support Thresholds

To further explore the performance of CE-EP, MB-EP, CAEP, CBA and CMAR, we conduct sensitivity analysis on the predictive accuracy and the number of selected patterns, of CE-EP, MB-EP, CAEP, CBA, and CMAR under seven minimum support threshold values in the following subsections.

TABLE 20 SENSITIVITY ANALYSIS ON PREDICTIVE ACCURACY (%): CE-EP, MB-EP, AND CAEP

Dataset	**CE-EP**			**MB-EP**			**CAEP**		
	max	**min**	**Δaccu**	**max**	**min**	**Δaccu**	**max**	**min**	**Δaccu**
australian	83.39	83.39	0	83.39	83.39	0	84.71	84.26	0. 45
breast-w	96.88	96.22	0.26	96.88	96.22	0.26	96.88	96.22	0.26
crx	82.21	82.21	0	82.21	82.21	0	84.85	84.41	0.44
cleve	84.83	84.48	0.35	82.41	82.07	0.34	85.52	82.76	2.76
diabetes	72.11	72.11	0	71.18	71.18	0	68.95	68.82	0.13
german	71.50	70.60	0.9	71.30	70.50	0.8	72.80	69.00	3.8
house-votes	96.82	96.82	0	96.82	96.82	0	90.91	90.45	0.46
hepatitis	85.33	85.33	0	85.33	85.33	0	86.00	84.00	2
horse-colic	85.83	85.28	0.55	83.33	83.06	0.27	84.72	84.17	0.55
hypothyroid	72.82	72.82	0	72.82	72.82	0	67.78	67.06	0.72
heart	83.70	82.59	1.1	82.96	82.96	0	85.93	83.33	2.6
infant	94.92	94.88	0.04	94.78	94.73	0.05	/	/	/
ionosphere	91.76	90.88	0.88	91.76	90.59	0.17	90.29	89.12	1.17
kr-vs-kp	92.23	91.70	0.53	91.54	90.47	1.07	83.49	81.70	1.79
labor	96.00	96.00	0	92.00	92.00	0	94.00	92.00	2
liver	61.76	61.76	0	61.76	61.76	0	57.65	57.65	0
mushroom	96.18	94.85	1.33	95.54	94.19	1.35	96.18	94.30	1.88
pima	72.11	72.11	0	71.18	71.18	0	68.82	68.82	0
promoters	72.00	72.00	0	72.00	72.00	0	/	/	/
spect	72.69	72.69	0	72.69	72.69	0	69.23	65.77	3.46
spectf	83.85	83.85	0	85.00	84.62	0.38	/	/	/
tictactoe	69.58	69.26	0.32	69.58	69.26	0.32	82.95	67.26	15.69
vote	95.95	95.95	0	95.95	95.95	0	90.00	89.05	0.95
wdbc	81.79	81.61	0.18	83.39	82.86	0.53	81.96	80.71	1.25
madelon	59.00	59.00	0	60.60	60.60	0	/	/	/
hiva	93.70	93.70	0	93.67	93.67	0	/	/	/
ovarian-cancer	92.86	92.38	0.48	92.86	92.38	0.48	/	/	/
lymphoma	77.73	77.27	0.46	77.73	77.27	0.46	/	/	/
dexter	88.33	88.33	0	89.33	89.00	0.33	/	/	/
arcene	86.67	86.67	0	86.67	86.67	0	/	/	/
breast-cancer	92.22	92.22	0	91.48	91.11	0.37	/	/	/
dorothea	95.06	95.06	0	95.06	95.06	0	/	/	/
colon	95.00	95.00	0	90.00	90.00	0	/	/	/
leukemia	100	100	0	100	100	0	/	/	/
lung-cancer	99.44	99.44	0	100	99.44	0.56	/	/	/
prostate	94.00	93.00	1	94.00	93.00	1	/	/	/

5.5.1 Sensitivity Analysis on CE-EP, MB-EP, and CAEP

Table 20 shows the change of predictive accuracy of CE-EP, MB-EP, and CAEP under seven support

thresholds, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, and 0.4. In the second row of Table 20, "max", "min", and "Δaccu" denote the maximum predictive accuracy under seven support thresholds, minimum predictive accuracy under seven support thresholds and the difference between the maximum predictive accuracy and minimum predictive accuracy. From Table 20, we can see that on the predictive accuracy, CE-EP and MB-EP are less sensitive to the support thresholds than CAEP. For all datasets, both CE-EP and MB-EP are insensitive to the different support thresholds, even for those high-dimensional datasets. Furthermore, CE-EP is not only more insensitive, but also always achieves higher accuracy under all the seven support thresholds than CAEP. In fact, from Table 20, on the 24 UCI datasets, CAEP is also insensitive to the support thresholds except for the tictactoe dataset.

TABLE 21 SENSITIVITY ANALYSIS ON THE NUMBER OF PATTERNS: CE-EP, MB-EP, AND CAEP

Dataset	CE-EP			MB-EP			CAEP		
	max	min	ratio	max	min	ratio	max	min	ratio
australian	51	27	1.9	51	27	1.9	1665	76	21.9
breast-w	449	171	2.6	449	171	2.7	449	171	2.6
crx	22	18	1.2	22	18	1.2	2066	77	26.8
cleve	56	33	1.7	71	40	1.8	892	73	12.2
diabetes	22	18	1.2	30	22	1.4	77	44	1.8
german	34	28	1.2	42	32	1.3	16649	136	122.4
house-votes	10	10	1.0	10	10	1.0	1616	330	4.9
hepatitis	19	17	1.1	24	14	1.7	1335	114	11.7
horse-colic	31	22	1.4	44	27	1.6	9383	306	30.7
hypothyroid	27	20	1.4	27	20	1.4	262	104	2.5
heart	43	28	1.5	74	36	2.1	599	72	8.3
infant	94	60	1.6	206	71	2.9	/	/	/
ionosphere	39	28	1.4	59	35	1.7	7125	503	14.2
kr-vs-kp	86	36	2.4	1037	93	11.2	6130	92	66.6
labor	38	37	1.0	53	52	1.0	2000	124	16.1
liver	4	4	1.0	4	4	1.0	27	24	1.13
mushroom	450	129	3.5	2076	216	9.6	2449	226	10.8
pima	22	18	1.2	27	22	1.2	77	34	2.3
promoters	8	8	1.0	8	8	1.0	/	/	/
spect	12	12	1.0	12	12	1.0	6832	88	77.6
spectf	27	23	1.2	149	67	2.2	/	/	/
tictactoe	56	30	1.9	56	30	1.9	641	54	11.9
vote	18	18	1.0	18	18	1.0	2938	554	5.3
wdbc	60	36	1.7	148	48	3.1	6132	600	10.2
madelon	22	22	1.0	33	28	1.2	/	/	/
hiva	31	28	1.1	206	71	2.9	/	/	/
ovarian-cancer	17	18	0.9	17	18	0.9	/	/	/
lymphoma	57	32	1.8	57	32	1.8	/	/	/
dexter	40	34	1.8	65	51	1.3	/	/	/
arcene	29	21	1.4	29	21	1.4	/	/	/
breast-cancer	51	40	1.3	276	93	3.0	/	/	/
dorothea	52	41	1.3	74	55	1.3	/	/	/
colon	27	25	1.1	57	56	1.0	/	/	/
leukemia	54	54	1.0	2416	1493	1.6	/	/	/
lung-cancer	39	38	1.0	55	51	1.1	/	/	/
prostate	134	74	1.8	134	74	1.8	/	/	/

Table 21 shows the change of the numbers of selected EPs of CE-EP, MB-EP, and CAEP under seven support thresholds. In the second row of Table 21, "max", "min", and "ratio" denote the maximum number of selected EPs under seven support thresholds, minimum number of selected EPs under seven support thresholds and the ratio of the maximum number of selected EPs and minimum number of selected EPs. As shown in Table 21, on the numbers of selected EPs, CE-EP and MB-EP are less sensitive to the support thresholds than CAEP. Both CE-EP and MB-EP are very insensitive to the different support

thresholds, even for those high-dimensional datasets, while CAEP is sensitive to the support thresholds on the 24 UCI datasets.

From Tables 20 to 21, we can conclude that on both low and high dimensional datasets, CE-EP and MB-EP are insensitive to the support thresholds on both predictive accuracy and the number of selected EPs. We can also come to the conclusion that EPs-based classifiers, CE-EP, MB-EP, and CAEP, are insensitive to the support thresholds on the predictive accuracy, although CE-EP and MB-EP are less sensitive to the support thresholds than CAEP. The explanation is that the EPs denote a strong contrast between classes, thus they have very strong differentiating power to predict each class. For example, although CAEP is sensitive to the support threshold on the number of selected EPs, this has little impact on its predictive accuracy.

5.5.2 Sensitivity Analysis on CE-EP, CBA, and CMAR

Since MB-EP has an extremely similar performance with CE-EP, in the following subsections, we only have the sensitivity analysis on CE-EP, CBA and CMAR under seven support thresholds, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, and 0.4. In Fig. 5, we plot the predictive accuracy of CE-EP, CBA and CMAR on the 24 UCI datasets in Table 6 with seven support thresholds.

As shown in Fig.5, CE-EP is less sensitive to the support thresholds than CBA and CMAR and always achieves a higher accuracy under all the seven support thresholds than CBA and CMAR on most datasets. The choice of the support thresholds is the key to both CBA and CMAR. For example, in Fig. 5, when the minimum support threshold is up to 0.3 or 0.4, the corresponding accuracies of both CBA and CMAR are greatly reduced.

On some datasets, the accuracies of CBA and CMAR are even reduced to 0, such as *clever*, *diabetes*, *german*, *ks-vr-kp*, *liver*, *pima*, *promoters*, *spect* and *tictactoe*. On the *infant-mortality*, *spectf*, and *wdbc* data sets, CBA doesn't work on all seven support thresholds due to the huge number of candidate patterns.

In Fig.6, we have a further sensitivity analysis of CE-EP, CBA, and CMAR on the number of selected patterns. We plot the numbers of selected patterns of CE-EP, CBA and CMAR on all 24 UCI datasets with the seven minimum support thresholds, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, and 0.4. From Fig.6, we can see that the numbers of selected patterns of CBA and CMAR are sensitive to the support thresholds. With a small support threshold, CBA and CMAR select a large number of association rules. With the support threshold increasing, the number of selected rules is greatly reduced.

Along with Fig. 5, we can conclude that when the support threshold is small, CBA and CMAR can obtain a large number of association rules for classification to achieve good accuracy as shown in Fig. 5. When the support threshold is large, CBA and CMAR prune too many rules, including useful rules, which results in low predictive accuracy. For example, in Fig.5, when the minimum support threshold

moves up to 0.4, on some datasets, the accuracies of CBA and CMAR are even reduced to 0, such as *clever*, diabetes, *german*, *ks-vr-kp*, *liver*, *pima*, *promoters*, *spect* and *tictactoe*. The explanation is that CBA and CMAR can no longer select any rules under this minimum support threshold, as shown in Fig.6. From Figures 5 and 6, it is clear that with a small support threshold, a large number of association rules provide rich information for classification and make CBA and CMAR achieve high accuracy. But in this case, it is also difficult to store, retrieve and maintain a large number of candidate patterns for classification. For example, in Fig.6, on the *infant-mortality*, *spectf*, and *wdbc* datasets, CBA doesn't work under a small support threshold due to the huge number of rules.

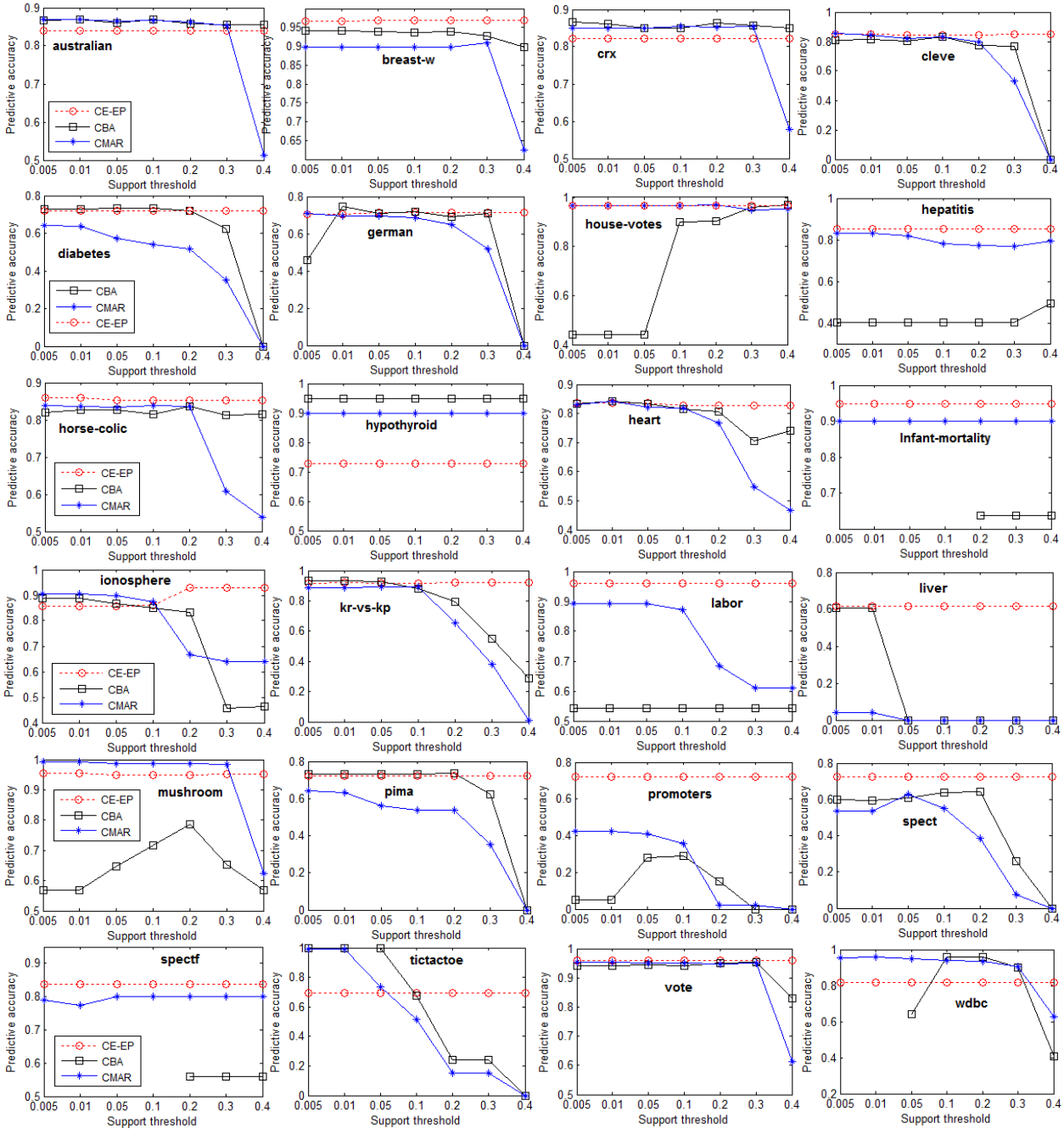

Figure 5: Sensitivity analysis on predictive accuracy: CE-EP, CBA, and CMAR

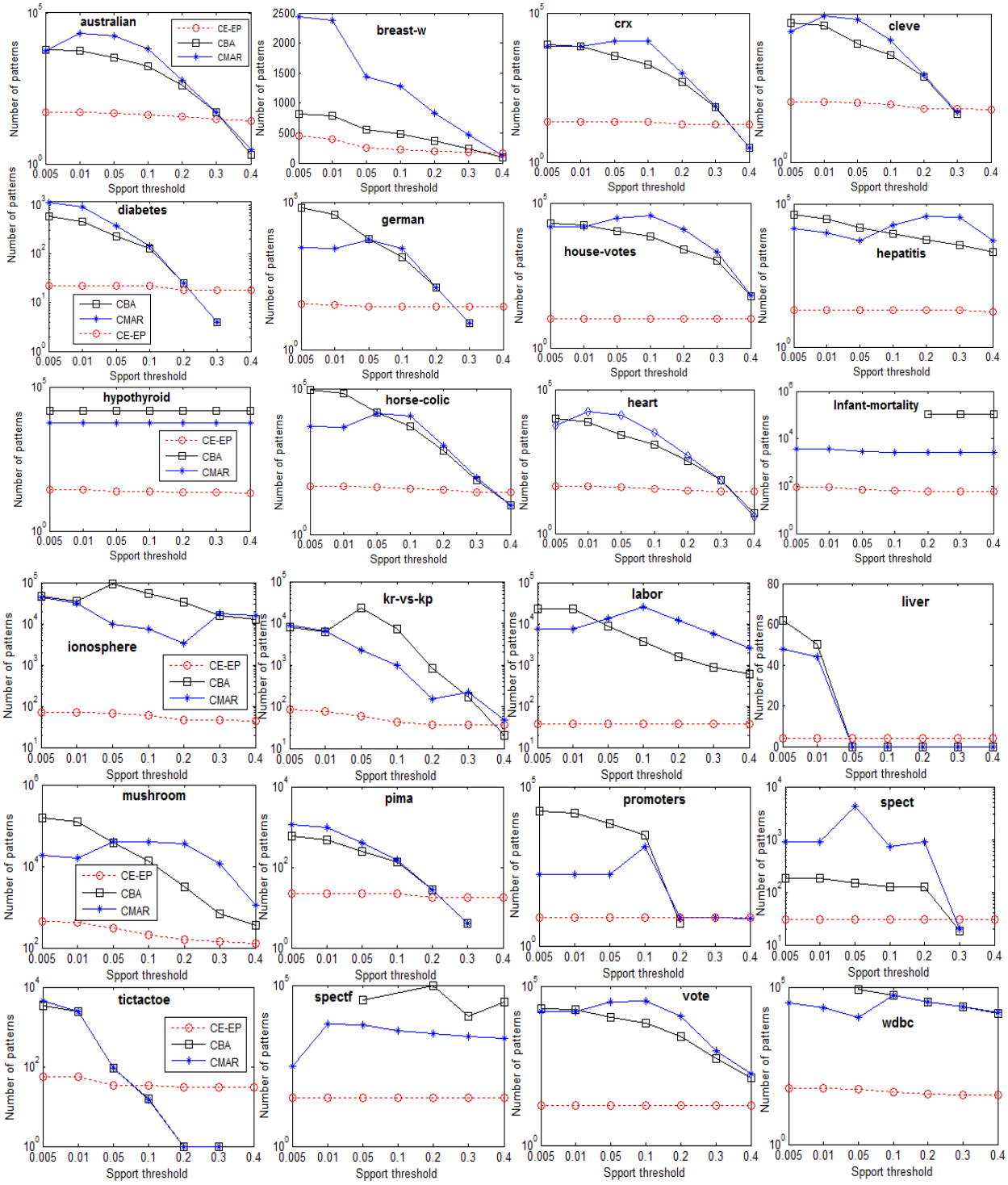

Figure 6: Sensitivity analysis on the number of patterns: CE-EP, CBA, and CMAR

As for CE-EP, under all seven support thresholds, it not only selects a small set of EPs, but also is insensitive to the support threshold. Therefore, these results further validate the theoretical analysis of the relationships between causal relevance and EP discriminability. More specifically, when CE-EP mines the EPs from the space of the direct causes and direct effects of the class attribute, they can achieve strongly predictive EPs no matter whether the support threshold is small or large. This also explains why

the accuracy of CE-EP remains stable under the different support thresholds as shown in Fig. 5.

5.6 Sensitivity Analysis on the Minimal Grow-rate thresholds

To explore the impact of the minimal grow-rate threshold on the performance of both CE-EP and CAEP, we conduct an analysis on the predictive accuracy of CE-EP and CAEP under seven minimum growth-rate thresholds, as shown in Figures 7 and 8, where GR stands for Growth Rate thresholds and the minimum support threshold is fixed at 0.1. Since MB-EP has an extreme similar performance with CE-EP, we don't plot the performance of MB-EP under the seven minimum growth-rate thresholds. On *infant*, *ionosphere*, *promoters* and *spectf*, CAEP cannot run under all seven growth-rate thresholds, and therefore Fig.7 plots the predictive accuracy of the other 20 low-dimensional datasets under the seven growth-rate thresholds. In Figure 8, the X-axis denotes all of the thirty six datasets corresponding to Table 6. Figures 7 to 8 show that both CAEP and CE-EP are not sensitive to the minimum growth-rate thresholds at all, especially CE-EP.

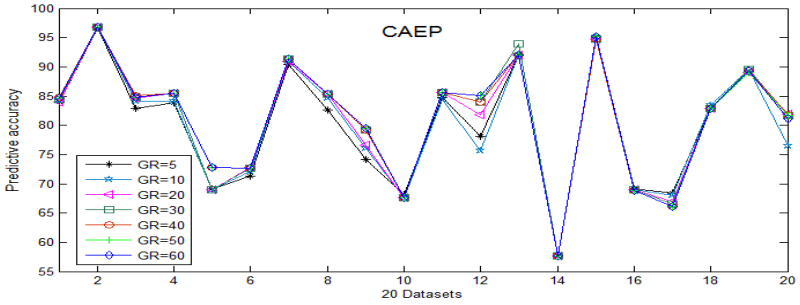

Figure 7: The impact of growth-rate thresholds on CAEP (The 20 datasets on the X-axis are: 1.australian, 2. breast-w, 3.crx, 4.cleve, 5.diabetes, 6.german,7. house-votes, 8.hepatitis, 9.horse-colic, 10. hypothyroid, 11.heart, 12.kr-vs-kp, 13.labor, 14. liver, 15.mushroom, 16. pima, 17.spect, 18.tictactoe, 19. vote, 20. wdbc).

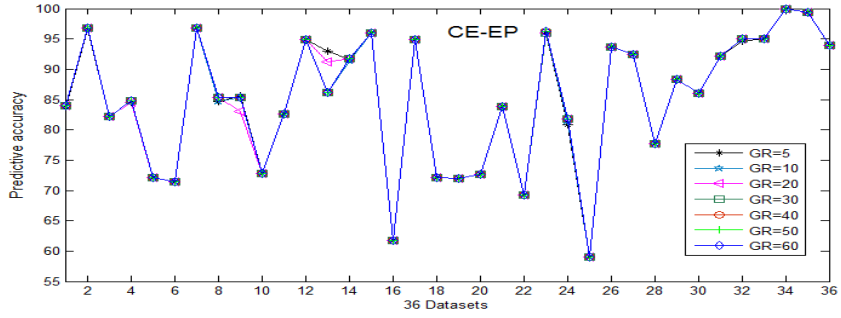

Figure 8: The impact of growth rate thresholds on CE-EP

5.7 Summary on the Experimental Results

Based on the comparative study in Sections 5.2 to 5.6, we have the following findings:

(1) With the seven support thresholds, the experiments demonstrate that the EPs-based classifiers are less sensitive to the support thresholds than the associative classifiers. Especially, on the predictive accuracy, EPs-based classifiers, CE-EP, MB-EP and CAEP, are less sensitive than associative classifiers, CBA and CMAR. As for the number of selected patterns, CE-EP and MB-EP are less sensitive than CAEP, CBA and CMAR, while CAEP is less sensitive than CBA and CMAR. Thus, the choice of a suitable support threshold is the key to control CBA and CMAR while it is not crucial for CE-EP, MB-EP and CAEP. Moreover, the study on the minimal grow-rate

thresholds verifies that both CE-EP and CAEP are insensitive to the minimal grow-rate thresholds on the predictive accuracy.

(2) The CE-EP and MB-EP classifiers are more accurate than the five associative classifiers (CAEP, SJEP, CBA, CMAR, and CPAR) and the five state-of-the-art non-associative classifiers (NB, Knn, J48, Bagging, and AdaBoost), and are very competitive with SVM. Meanwhile, both our classifiers produce smaller numbers of EPs. Moreover, CAEP, SJEP, CBA, CMAR and CPAR cannot handle high-dimensional datasets. As for the running time, CE-EP and MB-EP are faster than CAEP on all datasets. This verifies our theoretical analysis of the relationships between causal relevance and EP discriminability to help avoid generating non-EPs or redundant EPs in advance.

(3) Both the CE-EP and MB-EP classifiers can handle very high feature dimensions well yet get very promising performance. Although the MB-EP classifier considers the information of direct causes of the direct effects of the class attribute, it gets extremely similar performance with the CE-EP classifier. This validates that the direct causes and direct effects of the class attribute in causal Bayesian networks give a natural interpretation of what happens to the class attribute, and then naturally endows EPs with strong discriminating power.

6 Conclusions

How to mine EPs from high-dimensional data is a challenging issue in EP mining. Meanwhile, how to deal with high sensitivity to minimal support thresholds is another challenging problem. In this paper, we have brought causal relevance and EP discriminability together to reduce the pattern space of EP mining to the direct causes and direct effects or the Markov blanket of the class attribute in causal Bayesian networks, and proposed a new framework for building accurate EPs-based classifiers from high-dimensional data. Extensive experiments on a broad range of datasets have demonstrated the effectiveness of the proposed approach against other well-established methods, in terms of predictive accuracy, pattern numbers, running time, and sensitivity analysis on the minimal support thresholds and minimal grow-rate thresholds.

Acknowledgments

This work is supported by the National 863 Program of China (2012AA011005), the National 973 Program of China under grant 2013CB329604, the National Natural Science Foundation of China (61229301, 61070131, 61175051 and 61005007), the US National Science Foundation (CCF-0905337), and the US NASA Research Award (NNX09AK86G).

References

[1] C. F. Aliferis, I. Tsamardinos, A. Statnikov and L.E. Brown. (2003) Causal explorer: a causal probabilistic network learning toolkit for biomedical discovery. METMBS'03.

[2] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. Koutsoukos. (2010) Local causal and markov blanket induction for causal discovery and feature selection for classification part I: Algorithms and empirical evaluation. Journal of Machine Learning Research, 11, 171-234.

[3] E.Baralis, S.Chiusano, and P. Garza. (2008) A Lazy Approach to Associative Classification. IEEE Transactions on Knowledge and Data Engineering, 20(2): 156-171.

[4] C.L. Blake and C.J. Merz. (1998) UCI Repository of machine learning databases.

[5] A-L. Boulesteix, G. Tutz, and K. Strimmer. (2003) A CART-based approach to discover emerging patterns in microarray data. Bioinformatics, 19(18):2465–2472.

[6] C. P. de Campos and Q. Ji. (2011) Efficient structure learning of Bayesian networks using constraints. Journal of Machine Learning Research, 12, 663–689.

[7] C-L. Chen, F. S. C. Tseng and T. Liang. (2011)An integration of fuzzy association rules and WordNet for document clustering. Knowledge and Information Systems, 28(3), 687-708.

[8] D. M. Chickering, D. Heckerman, and C. Meek. (2004) Large-sample learning of Bayesian networks is NP-hard. Journal of Machine Learning Research, 5, 1287-1330.

[9] G. Dong and J. Li. (1999) Efficient mining of emerging patterns: discovering trends and differences. KDD'99, 43-52.

[10] G. Dong, X. Zhang, L. Wong, and J. Li. (1999) CAEP: classification by aggregating emerging patterns. DS'99, 30-42.

[11] H. Fan and K. Ramamohanarao. (2006) Fast discovery and the generalization of strong jumping emerging patterns for building compact and accurate classifiers. IEEE Transactions on Knowledge and Data Engineering 18(6), 721-737.

[12] G. Fang, G. Pandey, W. Wang, M. Gupta, M. Steinbach, and V. Kumar. (2012) Mining low-support discriminative patterns from dense and high-dimensional data. IEEE Transactions on Knowledge and Data Engineering, 24(2), 279 - 294.

[13] M. García-Borroto, J. Martínez-Trinidad and J. Carrasco-Ochoa. ((2011) Fuzzy emerging patterns for classifying hard domains. Knowledge and Information Systems, 28(2), 473-489.

[14] I. Guyon, C. F. Aliferis and A. Elisseeff. (2007) Causal feature selection in chapter of computational methods of feature selection, 63–86. Chapman and Hall.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. (2009) The WEKA data mining software: An Update. SIGKDD Explorations, 11(1).

[16] J.G. Lee, J. Han, X. Li, and H. Cheng. (2011) Mining Discriminative Patterns for Classifying Trajectories on Road Networks. IEEE Transactions on Knowledge and Data Engineering, 23(5): 713-726.

[17] J. Li, G. Dong, and K. Ramamohanarao. (2000) Making use of the most expressive jumping emerging patterns for classification. PAKDD'00, 220-232.

[18] J. Li, G. Dong and K. Ramamohanarao (2000). Instance-based classification by emerging patterns. PKDD'00, 191-200.

[19] W. Li, J. Han and J. Pei. (2001) CMAR: accurate and efficient classification based on multiple-class association rule. ICDM'01, 369–376.

[20] J. Li, H. Liu, J.R. Downing, A.E.J. Yeoh and L. Wong. (2003) Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (all) patients. Bioinformatics, 19(1), 71–78.

[21] J. Li and L. Wong. (2002) Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. Bioinformatics, 18(5), 725-734.

[22] B. Liu, W. Hsu and Y. Ma. (1998) Integrating classification and association rule mining. KDD'98, 80-86.

[23] H. Liu, Y. Lin and J. Han. (2011) Methods for mining frequent items in data streams: an overview. Knowledge and Information Systems, 26(1), 1-30.

[24] D. Lo, H. Cheng, J. Han, S. Khoo, and C. Sun. (2009) Classification of software behaviors for failure detection: a discriminative pattern mining approach. KDD'09, 557-566.

[25] E. Loekito and J. Bailey. (2006) Fast mining of high dimensional expressive contrast patterns using zero suppressed binary decision diagrams. KDD'06, 307–316.

[26] LUCS KDD Software Library. (2012) http://www.csc.liv.ac.uk/~frans/KDD/Software.

[27] D. Margaritis and S. Thrun. (2000) Bayesian Network induction via local neighborhoods. NIPS'99 505-511, MIT Press,

Cambridge, MA.

[28] P. Novak, N. Lavrac, and G. Webb. (2009) Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining. Journal of Machine Learning Research, 10:377-403.

[29] J. Pearl. (1991) Probabilistic reasoning in intelligent systems. Morgan Kaufmann, San Francisco, California, 2nd edition.

[30] J.M. Peña, R. Nilsson, J. Björkegren, and J. Tegnér. (2007) Towards scalable and data efficient learning of Markov boundaries. International Journal of Approximate Reasoning, 45(2), 211-232.

[31] P. Spirtes, C. Glymour and R. Scheines. (2000) Causation, prediction, and search (Second ed.). The MIT press.

[32] B. Qin, Y. Xia and S. Prabhakar. (2011) Rule induction for uncertain data. Knowledge and Information Systems, 29(1), 103-130.

[33] A. Rodríguez-González, J. Martínez-Trinidad, J.Carrasco-Ochoa and J. Ruiz-Shulcloper (2011) RP-Miner: a relaxed prune algorithm for frequent similar pattern mining. Knowledge and Information Systems, 27(3), 451-471.

[34] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. Machine Learning, 65, 31–78.

[35] J. Wang and G. Karypis. (2005) HARMONY: efficiently mining the best rules for classification. SDM'05, 205-216.

[36] X. Yin and J. Han. (2003) CPAR: classification based on predictive association rule. SDM'03, 369–376.

[37] X. Zhang, G. Dong and K. Ramamohanarao. (2000) Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets. KDD'00, 310-314.

[38] X. Zhang, G. Dong, and K. Ramamohanarao. (2000) Information-based classification by aggregating emerging patterns. IDEAL'00, 48-53.

Kui Yu received the MSc degree in Computer Science from the Hefei University of Technology, China, in 2007. He is currently a Ph.D. student in the School of Computer Science and Information Engineering at the Hefei University of Technology (China). His research interests include feature selection, probabilistic graphical models and machine learning.

Hao Wang received his Ph.D. degree in Computer Science from the Hefei University of Technology, China, in 1997. He is a Professor of the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China. His research interests include robotics, artificial intelligence, data mining, probabilistic graphical models and machine learning.

Wei Ding received her Ph.D. degree in Computer Science from the University of Houston in 2008. She has been an Assistant Professor of Computer Science in the University of Massachusetts Boston since 2008. Her research interests include data mining, machine learning, artificial intelligence, computational semantics, and with applications to astronomy, geosciences, and environmental sciences. She has published more than 60 referred research papers, 1 book, and has 1 patent. She is an Associate Editor of Knowledge and Information Systems (KAIS) and an editorial board member of the Journal of System Education (JISE).

Dr. Ding is the recipient of a Best Paper Award at the 2011 IEEE International Conference on Tools with Artificial Intelligence (ICTAI), a Best Paper Award at the 2010 IEEE International Conference on Cognitive Informatics (ICCI), a Best Poster Presentation award at the 2008 ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPAITAL GIS), and a Best PhD Work Award between 2007 and 2010 from the University of Houston. Her research projects are currently sponsored by NASA and DOE.

Xindong Wu is a Yangtze River Scholar in the School of Computer Science and Information Engineering at the Hefei University of Technology (China), a Professor of Computer Science at the University of Vermont (USA), and a Fellow of the IEEE. He received his Bachelor's and Master's degrees in Computer Science from the Hefei University of Technology, China, and his Ph.D. degree in Artificial Intelligence from the University of Edinburgh, Britain. His research interests include data mining, knowledge-based systems, and Web information exploration.

Dr. Wu is the Steering Committee Chair of the IEEE International Conference on Data Mining (ICDM), the Editor-in-Chief of Knowledge and Information Systems (KAIS, by Springer), and a Series Editor of the Springer Book Series on Advanced Information and Knowledge Processing (AI&KP). He was the Editor-in-Chief of the IEEE Transactions on Knowledge and Data Engineering (TKDE, by the IEEE Computer Society) between 2005 and 2008. He served as Program Committee Chair/Co-Chair for ICDM '03 (the 2003 IEEE International Conference on Data Mining), KDD-07 (the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining), and CIKM 2010 (the 19th ACM Conference on Information and Knowledge Management).