Causal Associative Classification

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

[1]Department of Computer Science
Hefei University of Technology
Hefei, 230009, China

[2]Department of Computer Science
University of Vermont
Burlington, VT 05405, USA

[3]Department of Computer Science
University of Massachusetts
Boston, MA 02125, USA

ykui713@gmail.com; xwu@cs.uvm.edu; ding@cs.umb.edu; jsjxwangh@hfut.edu.cn; lhy_y@sohu.com

Abstract—Associative classifiers have received considerable attention due to their easy to understand models and promising performance. However, with a high dimensional dataset, associative classifiers inevitably face two challenges: (1) how to extract a minimal set of strong predictive rules from an explosive number of generated association rules; and (2) how to deal with the highly sensitive choice of the minimal support threshold. In order to address these two challenges, we introduce causality into associative classification, and propose a new framework of causal associative classification. In this framework, we use causal Bayesian networks to bridge irrelevant and redundant features with irrelevant and redundant rules in associative classification. Without loss of prediction power, the feature space involved with the antecedent of a classification rule is reduced to the space of the direct causes, direct effects, and direct causes of the direct effects, a.k.a. the Markov blanket, of the consequent of the rule in causal Bayesian networks. The proposed framework is instantiated via baseline classifiers using emerging patterns. Experimental results show that our framework significantly reduces the model complexity while outperforming the other state-of-the-art algorithms.

Keywords- associative classification; emerging patterns; causal bayesian networks

I. INTRODUCTION

Associative classification integrates association rule mining and 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 good results and yield an interpretable model. Liu et al. introduced CBA (Classification Based on Associations), the first associative classifier [16]. Later, Dong and Li discussed a new type of association patterns named emerging patterns [10]. An emerging pattern is a combination of attribute values that occurs mostly in one class but barely appears in the remaining classes; so the presence of such a pattern in a query object gives some evidence about the class to which the object should belong. Dong et al. proposed the first associative classifier based on emerging patterns, called CAEP (Classification by Aggregating Emerging Patterns) [11]. Associative classification using emerging patterns has shown to be a powerful method for constructing accurate classifiers, even for imbalanced data [12, 19]. Although many research efforts have significantly advanced the techniques for associative classification, two challenging issues remain.

(1) How to extract a minimal set of strong predictive rules from an explosive number of generated association rules?

Most associative classifiers are constructed in two steps: generating frequent patterns satisfying certain minimum support and confidence constraints, and then making predictions based on the selected patterns. Although many pruning strategies have been proposed, a huge number of rules can still be discovered from high dimensional data even using a rather high minimum support threshold. It is difficult to store, retrieve, prune, and sort a large number of rules efficiently for classification; it hampers the understanding of the final classifier; and it leads to overfitting. These problems exist mainly because the existing associative classifiers only focus on generating rules with the support-confidence framework, without considering the predictive ability of the features involved in a classification rule. Thus, extracting a minimal set of rules with strongly predictive ability is critical to build an efficient associative classifier from high dimensional data which are prevalent in many real-world applications.

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

The performance of most existing associative classifiers significantly depends on the support threshold value. For a large dataset, a small support value could generate a huge amount of rules while a large value might cause serious accuracy degradation. Thus it is a daunting problem to find an appropriate support threshold value.

To battle these challenges, we propose a new framework to introduce causality into associative classification. Traditional associative classification algorithms identify relationships between the class label and its antecedents using statistical correlation. However, “correlation is not causation.” [20]. For example, the growth of hair can be strongly correlated with the growth of fingernails in a given database, but one of them does not cause the other. In this case, the correlation only reveals a spurious association, due to a possibly unobserved common cause. Consequently, by detecting causal relationships between the class label and its antecedents, we can uncover causal or consequential factors with respect to the class label.

The causal or consequential factors of a variable give the natural interpretation of most of the events occurring in real-world applications. We use causal Bayesian networks to bridge irrelevant and redundant features with irrelevant and redundant rules in associative classification. Without loss of prediction power, the feature space involved with the antecedent of a classification rule is reduced to the space of the direct causes, direct effects, and direct causes of the direct effects, a.k.a. Markov blanket, of the consequent of this rule in causal Bayesian networks.

When generating the set of classification rules, the only features taken into account are those that belong to this causal feature space instead of the combinations of all features. Thus, it can greatly reduce computational cost and large resource demand in the stage of rule extraction. Furthermore, the extracted rules are not only causally interpretable, but also causally informative. The proposed framework is instantiated via baseline classifiers using emerging patterns (EP for short) with two new classifiers, CE-EP and MB-EP classifiers (with CE for direct **C**auses and direct **E**ffects and MB for **M**arket **B**lanket, respectively).

The reminder of the paper is structured as follows. Section 2 reviews related work on associative classification. Section 3 provides a theoretical analysis on how to extract rules with causality, and then presents a causal associative classification framework. Experimental results are reported in Section 4, and we conclude in Section 5.

II. Related Work

Associative classification is a scientific study which integrates association rule discovery and classification to a model for prediction. Existing associative classifiers can be grouped into two categories. One is based on general association rules and the other is based on emerging patterns. Successful algorithms in the first category include CBA [16], CMAR [15] and CPAR [25]. 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 it selects the best rule to be applied to each test instance [16]. Based on CBA, Li et al. [15] 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 [25]. 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.

A new type of association pattern, called emerging patterns, was first introduced by Dong and Li [10]. It captures emerging trends in time-stamped datasets, or useful contrasts between data classes. An emerging pattern is a multivariate pattern whose support value increases sharply from a background dataset to a target dataset. Emerging patterns have shown very successful results on constructing accurate and robust classifiers. Compared to the approach based on general association rules, associative classification based on emerging patterns uses the aggregation of many groups of emerging patterns while the other approach uses general association rules. For example, CBA uses one association rule group at a time.

Much research on emerging patterns has largely focused on the use of the discovered patterns for classification, such as classification by emerging patterns, CAEP classifier and its variants [11, 17] and classification by jumping emerging patterns [11, 14]. A jumping emerging pattern is a special emerging pattern whose support increases from zero in one dataset class to non-zero in the other dataset class. In addition, a strong jumping emerging pattern is a special jumping emerging pattern whose support is zero in one dataset class but non-zero, and satisfying a given minimal threshold in the other dataset class [12]. Emerging pattern classifiers have been proved to be very valuable tools to solve real problems in many fields [7].

Causal discovery has found wide applications in science and technology. Since late 1980's, the work on formal theories of causality and causal induction by Spirtes, Glymour, Pearl, Cooper and others has been gaining ground [18, 8, 20]. A Bayesian network, as one of the most common models to represent causal relationships between features, has been extensively studied. But learning a full Bayesian network from observational data is an NP-hard problem [9]. Consequently, local causal induction which has extended learning Bayesian Networks with tens of thousands of variables has received considerable attention [23, 3]. We adopt this local approach in this paper. The importance of introducing causality into the machine learning and data mining community has been well acknowledged [20, 13, 6]. In this paper, we introduce causality into associative classification and propose a new framework of associative classification with causality.

III. Causal Associative Classification

A. *Extracting Rules with Causality*

Causal discovery is in the center of human reasoning and decision-making. Because causal Bayesian networks provide a convenient framework for reasoning about causality among random variables, to simplify our presentation, we limit ourselves to causal Bayesian networks to represent causal relationships between variables in this paper. We will denote variables in uppercase letters X; Y; S, values in lowercase letters, x; y; s, respectively. A target (i.e., the class label) variable is denoted as C unless stated otherwise. The words "variable," "node", and "feature" are all used interchangeably in the rest of this paper.

A Bayesian network is a mathematical tool that compactly represents a joint probability distribution P among a set of random variables V using a directed acyclic graph G, which is annotated with conditional probability tables of the probability distribution of a variable given any instantiation of its parents. We call the triplet <V, G, P> a Bayesian network [18]. A simple Bayesian network for lung cancer is shown in Fig. 1 [13]. The states of the nodes and the probabilities that are associated with this structure are not shown for better clarify.

A causal Bayesian network <V, G, P> is a Bayesian network with the additional semantics that $X \in V$ and $Y \in V$

if a node X is a parent of a node Y in G, then X directly causes Y [21]. For example, Fig.1 can also be treated as an example of a causal Bayesian network.

Definition 1 [18] **(Conditional Independence)** Two variables $X \in V$ and $Y \in V$ are said to be conditionally independent given some set of variables $S \subseteq V$ if, for any assignment of values x, y, and s to the variables X, Y, and S respectively, P(X=x | S=s,Y=y)=P(X =x|S= s) holds.

Definition 2 [18] **(Causal Markov Condition)** In a causal Bayesian network, if every node is independent of its non-effects given its direct causes, then the causal Markov condition holds.

Definition 3 [18] **(Faithfulness)** A Bayesian network G and a joint distribution P are faithful to one another if and only if every conditional independence entailed by the graph of G is also present in P.

Definition 4 [18] **(Causal faithfulness)** A causal Bayesian network is faithful if it satisfies the faithfulness condition of Definition 3.

Let the node "Lung Cancer" be a target node in Fig. 1. Causal relationships between "Lung Cancer" and the rest nodes are classified into three categories, irrelevant nodes, Markov blanket, and redundant nodes as follows.

Definition 5 (Irrelevant Nodes) In a causal Bayesian network, if node X is disconnected from Y, then X is a irrelevant node with respect to Y.

If a node X is irrelevant to Y, it cannot carry any information about Y at all, no matter what the context is. For example, the node "Born an Even Day" in Fig. 1 is disconnected from "Lung Cancer," thus the rule: *{"Born an Even Day"=yes}* or *{"Born an Even Day"=no}* cannot provide any predictive information to the target node of Lung Cancer.

Definition 6 [18, 21] **(Markov Blanket)** In a causal Bayesian network with causal faithfulness, the Markov blanket of the node Y, denoted as MB(Y), is the set of its direct causes, direct effects and direct causes of the direct effects.

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 causes of the direct effects: "Allergy".

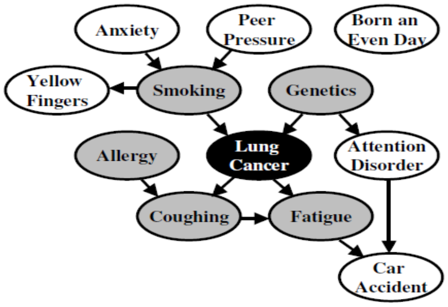


Figure 1. A causal Bayesian network for lung cancer

Definition 7 (Redundant Nodes) In a causal Bayesian network, if a node X has a connecting path to Y and doesn't belong to MB(Y), then it is a redundant node with respect to Y.

We speak of a redundant node with respect to the node Y, if its value is fully determined by a set of nodes that does not include Y. For example, in Fig.1, according to causal Markov condition (see Definition 2), once all the direct causes of node Lung Cancer have been given, the value of an indirect cause (e.g. "Anxiety") is fully determined by the direct causes of node Lung Cancer and does not bring any additional information to the node. For instance, increased "Anxiety" will increase "Smoking," but cannot influence directly "Lung Cancer," when the value of "Smoking" is known already. Consequently, with two rules: *{"Smoking"=yes}→{"Lung Cancer"=yes}* and *{"Anxiety"=yes and "Smoking"=yes}→{"Lung cancer"=yes}*, it suffices to have *{"Smoking"=yes}→ {"Lung Cancer"=yes}* as a predictive rule, and we do not need to know about "Anxiety."

Property 1 [21] In causal Bayesian networks with causal faithfulness, the MB(Y) is satisfied with the following property:

$$\forall S \in V \setminus \{MB(Y), Y\}, P(Y \mid MB(Y), S) = P(Y \mid MB(Y)).$$

This property says that the Markov blanket of each node is the set of all nodes that are dependent of it, conditioned on all other nodes. In other words, in causal Bayesian networks, the direct causes, direct effects, and direct causes of the direct effects of Y store information about Y that cannot be obtained from any other nodes. For example, in Fig.1, if we know MB("Lung Cancer"), any nodes outside of the MB("Lung Cancer") would be redundant. Thus, it should be noted that if the antecedent of a classification rule includes both irrelevant and redundant nodes, not only many irrelevant and redundant rules are generated, but also the learning performance will be degraded.

The above observations motivate us to integrate causal induction with associative classification to address the two challenges on the minimal rule set and the impact of the minimal support threshold. We plan to explore this integration with two views:

(1) How to bridge the gap between irrelevant and redundant nodes in causal Bayesian networks and irrelevant and redundant classification rules in associative classification?

(2) How to reduce the search space of the antecedent of a classification rule in associative classification to the space of direct causes and effects or the Markov blanket of the consequent of this rule in causal Bayesian networks?

Before we address those two views in detail, we first give the definition of the j-measure which is defined in Eq. (1) as follows [5].

$$j(X;y) = \sum_x p(x \mid y)\log\left(\frac{p(x \mid y)}{p(x)}\right) \qquad (1)$$

The j-measure gives the information measure that the value of y of a random variable Y provides information about another random variable X whenever it changes the probabilities of the various possible (discrete) values {x} of X from their prior values {P(x)} to new, posterior values {P(x|y)}.

It was introduced into the rule induction literature by Smyth and Goodman [22] as an information theoretic means of quantifying the information content of a rule that is soundly based on theory. Given a rule of the form "if Y=y, then X=x", using the notation of [22], the (average) information content of the rule, measured in bits of information, can be calculated by Eq.(1). It is assumed that a rule with a high j-value will tend to have a high predictive accuracy.

Property 2 [5] **(Non-Negativity)** j(X,Y=y)⩾0.

Property 2 shows j(X,Y=y)=0 if and only if the prior values {P(x)} and posterior values {P(x|y)} are identical. From the viewpoint of rule induction, only when the rule {Y=y} provides no information about X at all, the term j(X,Y=y)=0 exists. Thus, we obtain Property 3.

Property 3 If j(X,Y=y)=0, then the rule{Y=y} cannot provide any predictive information to the variable X.

According to Property 3, we get the definition of an irrelevant rule as follows.

Definition 8 (Irrelevant Rules) Let the rule {X=x} be an arbitrary rule with respect to C. If the term $\forall S \subseteq V \setminus \{X, C\},\ j(C; X = x \mid S = s) = 0$ exists, then the rule {X=x} is an irrelevant rule with respect to C.

Definition 8 shows that an irrelevant rule cannot convey any further information to C with the presence of other rules. For example, with two rules A: *{S=s}→{C=c}* and B: *{X=x and S=s}→{C=c}* where *{S=s}* is an arbitrary rule set for C, if A conveys exactly the same predictive information as B, then *{X=x}* is an irrelevant rule.

Lemma 1 If X is an irrelevant node to the target node C in causal Bayesian networks, then with the presence of any other rules, the rule {X=x} is an irrelevant classification rule with respect to C in associative classification.

Proof: Since X is irrelevant to C in a causal Bayesian network, according to Definition 5, it cannot carry any information about C at all, no matter what the context is. Thus, X∈ V irrelevant to C∈ V can be formulated as Eq. (2):

$$\forall S \subseteq V \setminus \{X, C\} : P(C = c \mid S = s, X = x) = P(C = c \mid S = s). \quad (2)$$

Then the amount of information that the rule {X=x} conditioned on any other rules {S=s} gives about the target C is calculated below.

$$\begin{aligned} j(C; X = x \mid S = s) &= \sum_{c} p(c \mid x, s) \log\left(\frac{p(c \mid x, s)}{p(c \mid s)}\right) \\ &= p(c \mid x, s) \log\left(\frac{p(c \mid x, s)}{p(c \mid s)}\right) + p(\bar{c} \mid y) \log\left(\frac{p(\bar{c} \mid x, s)}{p(\bar{c} \mid s)}\right) \\ &= p(c \mid x, s) \log\left(\frac{p(c \mid x, s)}{p(c \mid s)}\right) + (1 - p(c \mid x, s)) \log\left(\frac{1 - p(c \mid x, s)}{1 - p(c \mid s)}\right) \\ &= 0 \ (according to\ Eq.(2)) \end{aligned}$$

Thus, from Definition 8, we get Lemma 1. □

For instance, in Fig.1, the node of "Born an Even Day" is an irrelevant node to the node of "Lung Cancer." It is evident that the rule: *{"Born an Even Day"=yes and "Smoking"=yes}→{"Lung Cancer"=yes}* provides the exactly same predictive information as the rule: *{"Smoking"=yes}→{"Lung Cancer"=yes}*.

Definition 9 (Redundant Rules) Let the rule {X=x} be an arbitrary rule with respect to C. If the term $\exists S \subseteq V \setminus C \text{ and } X \notin S,\ j(C; X = x \mid S = s) = 0$ holds, then it is a redundant rule with respect to C.

Definition 9 shows that with the presence of a rule set {S=s}, if the information of a rule {X=x} conveyed to C is fully contained by the rule set {S=s} that does not include X and C, then {X=x} is a redundant rule to C. For example, with rule A: *{X=x and S=s}→{C=c}*, if rule B: *{S=s}→{C=c}* contains exactly the same predictive information as A, then {X=x} is a redundant rule and A can be replaced by B.

Lemma 2 If a node X is a redundant node to the target node C in causal Bayesian networks, then the rule {X=x} is a redundant rule with respect to C in associative classification.

Proof: According to Definition 7 and Property 1, since the node X is a redundant node to the target C in causal Bayesian networks, there must have a subset $S \subseteq MB(C)$ for which there exist some values s and c to fully determine the value of X. By Property 1, we can formulate this as Eq.(3):
$\exists S \subseteq MB(C), X \notin MB(C)$

$$p(C = c \mid X = x, S = s) = p(C = c \mid S = s). \quad (3)$$

Thus, similar to the proof of the Lemma 1, with Eq.(3), we can get the term j(C=c,X=x|S=s)=0 where $S \subseteq MB(C), X \notin MB(C)$. That is to say, with the presence of a rule set of *{S=s}→{C=c}*, the rule *{X=x}* cannot provide any further information to the rule set of *{S=s}→{C=c}* where $S \subseteq MB(C), X \notin MB(C)$. From Definition 9, we obtain Lemma 2. □

For example, in Fig.1, given the value of node of "Smoking", the node of "Yellow Finger" is redundant to the node "Lung Cancer". Therefore, with two rules: *{"Smoking"=yes}→{"Lung Cancer"=yes}* and *{"Smoking"=yes and "Yellow Finger"=yes}→{"Lung cancer"=yes}*, according to causal Markov condition (see Definition 2), it is clear that *{"Smoking"=yes}* contains the information that conveyed by *{"Yellow Finger"=yes}* to the node of "Lung Cancer" . With the presence of the rule of *{"Smoking"=yes}*, *{"Yellow Finger"=yes}* is redundant. Thus, it suffices to have *{"Smoking"=yes}→{"Lung Cancer"=yes}* as a predictive rule.

Corollary 1 The search space of classification rules in associative classification can be reduced to the space of the Markov blanket of the class label in causal Bayesian networks.

Proof: With the results of Lemmas 1 and 2, we can remove irrelevant and redundant nodes with respect to the class label in causal Bayesian networks to achieve the goal of pruning irrelevant and redundant rules in associative classification. Thus, by removing irrelevant and redundant nodes in causal Bayesian networks, we can reduce the search space of highly predictive rules in associative classification to the space of Markov blanket of the class label in a causal Bayesian network. □

With Corollary 1, within the Markov blanket of the class label, both the direct causes and the direct effects have a direct connecting path to the class label while the direct causes of the direct effects (spouses) of the class label don't have. Thus, in causal Bayesian networks, the spouses cannot individually predicate the class label and only may enhance the predictive power of their direct effects when they join with its direct effects, and also, direct causes joint with direct effects give the most highly predictive ability to predict the class label. Therefore, the search space of classification rules in associative classification can be further reduced to direct causes and direct effects of the class label in causal Bayesian networks, and then we obtain Corollary 2 as follows.

Corollary 2 The search space of classification rules in associative classification can be further reduced to direct causes and direct effects of the class label in causal Bayesian networks.

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 predicate lung cancer. It may enhance the predictive power of “Coughing” when it joins with “Coughing.”

B. *Causal Associative Classification*

Table I introduces a new framework for causal associative classification. Using Lemmas 1 and 2 and corollaries 1 and 2, we introduce causality into associative classification.

TABLE I. CAUSAL ASSOCIATIVE CLASSIFICATION FRAMEWORK

1. Initialization

The feature set V, the class label C, the support and confidence threshold values

2. Causal induction phase

Remove irrelevant and redundant nodes and get direct causes and direct effects or Markov blanket of C.

3. Rule mining phase

(a) **MB (Markov Blanket) Approach:** Extract causal rules from the feature space of the Markov blanket of C by Corollary 1

Or

(b) **CE (direct Causes and direct Effects) Approach:** Extract causal rules from the feature space of direct causes and direct effects of C by Corollary 2

4. Building a classifier using extracted causal rules

The key tasks of our framework are (1) removing the causally irrelevant and redundant nodes to a target, and then identifying the direct causes and direct effects or the Markov blanket of a target, and (2) constructing causally informative classification rules.

Learning a full Bayesian network is a common method to address the first task. But it is an NP-hard problem. Fortunately, in associative classification, we only need to consider a rule if its consequent is the class label. Therefore, with the Corollaries 1 and 2, rather than to recover a whole causal Bayesian network among all variables, in fact, our framework only needs to uncover cause-effect relationships between a target variable and the rest or to uncover the Markov blanket of a target. Moreover, much research efforts have significantly advanced the local causal induction techniques, such as the HITON_MB and MMMB algorithms [3], which discover local cause-effect relationships around a target variable of interest in the form of direct causes and effects or Markov blankets.

Our framework can extract causal rules from this casual feature space, direct causes and direct effects or the Markov blanket of the class label. Thus, reducing the search space of the antecedent of a rule to this causal feature space naturally endows classification association rules with strongly causal interpretability, and those rules are also causally informative. Most importantly, in a causal Bayesian network with causal faithfulness, the direct causes and direct effects or the Markov blanket of a target is unique and minimal [18], and our framework has a good chance to generate a set of causal rules that is as small as possible. For example, in Fig.1, the direct causes and direct effects of lung cancer are the nodes “smoking”, “genetics”, “coughing” and “fatigue.” The rules extracted from the four variables naturally have strongly causally interpretable and causally informative. Consequently, the rule *{“Smoking”=yes, “Coughing”=yes}* → *{“Lung Cancer”=yes}* has a strongly causal ability to predict whether a person suffers from the disease of lung cancer. With a small set of causally informative and highly predictive rules, our framework should be less sensitive to the support threshold value.

IV. EXPERIMENTAL EVALUATION

A. *Experimental setup*

In order to thoroughly evaluate the proposed causal associative classifiers, twenty datasets (Table II) are selected from the UCI datasets, the WCCI2006 datasets and the NIPS2003 datasets, respectively. Of the twenty datasets, six are very high-dimensional datasets, including *madelon*, *hiva*, *ovarian-cancer*, *lymphoma*, *breast-cancer*, and *dorothea*. All datasets are in binary classes. In our experiments, we use classifiers based on emerging patterns as a baseline to instantiate our framework, that is, generating emerging patterns according to the CE and MB approaches described in Table I, denoted as CE-EP and MB-EP classifiers with CE standing for direct **C**auses and direct **E**ffects, MB for **M**arket **B**lanket, and EP for **E**merging **P**atterns, respectively.

TABLE II. SUMMARY OF DATASETS. #: THE NUMBER OF FEATURES , SIZE: THE NUMBER OF INSTANCES.

Datasets	#	SIZE	Datasets	#	SIZE
australia	14	690	spect	22	267
breast-w	9	3146	spectf	44	267
heart	13	270	vote	16	435
infant-mortality	86	5337	wdbc	30	569
ionosphere	34	351	madelon	500	2000
kr-vs-kp	36	3196	hiva	1617	4229
liver	6	345	ovarian-cancer	2190	216
mushroom	22	8124	lymphoma	7399	227

pima	8	768	breast-cancer	17816	286
promoters	57	106	dorothea	100000	800

Our evaluation is based on a comparison against the well-known associative classifiers CBA [16], CMAR [15], CPAR [25], CAEP [11] and Strong Jumping EP (SJEP for short) classifiers [12]. We also compare the predictive accuracy of our classifiers with the state-of-the-art non-associative classifiers, such as Naïve Bayes (NB), decision tree J48, Bagging and Boosting using their Weka implementation with the default parameters. In all of our experiments, we use 10-fold cross-validation on all datasets except for the datasets of *spect* and *ionosphere* which provide test datasets in addition to training datasets.

We use the method proposed by Aliferis et al. [2] to discretize continuous attributes. In the experiments, we set the minimum confidence threshold to 0.8 for CBA and CMAR, and set the growth rate to 20 for CAPE and our 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 EP classifier based on strong jumping EP uses the minimal support threshold suggested in the original paper [13]. The parameters for CPAR are set the same as those reported in [25]. CBA, CMAR, and CPAR are implemented in the LOCS KDD software [24] in Java, while CE-EP, MB-EP and CAEP are implemented in C++. The experiments are performed on a Window-XP PC with an Intel Pentium 2.5 GHz processor and 4.0GB RAM.

B. *Comparison of the classification accuracy of our two classifiers with the other five associative classifiers*

Table III reports the classification accuracy of the seven classifiers on the twenty benchmark datasets. The best results are highlighted in bold and the symbol "/" denotes either the classifier runs out of memory or it doesn't work due to a huge number of rules. Table IV summarizes win/tie/loss counts for comparing our two classifiers with the other classifiers. As shown in Tables III and IV, our two classifiers, CE-EP and MB-EP, usually outperform CAEP, SJEP, CBA, CMAR, and CPAR. Although CE-EP only considers mining the classification rules from the space of the direct causes and direct effects of the class label, it outperforms not only MB-EP, but also SJEP, CBA, CMAR, and CPAR. Empirical results indicate that detecting cause–effect relationships can find high quality predictive rules to achieve better accuracy. Most importantly, using causality to guide pattern generation can efficiently deal with high-dimensional datasets.

TABLE III. COMPARISON OF CLASSIFICATION ACCURACY (%) WITH SEVEN ASSOCIATIVE CLASSIFIERS

Datasets	**CE-EP**	**MB-EP**	**CAEP**	**SJEP**	**CBA**	**CMAR**	**CPAR**
australia	**66.47**	60.88	62.35	52.50	66.38	46.23	46.96
breast-w	**96.88**	**96.88**	**96.88**	90.80	94.09	90.82	92.95
heart	**83.33**	**83.33**	82.22	70.37	80.74	81.11	79.63
infant_mortality	**94.88**	94.78	/	/	63.72	90.00	84.30
ionosphere	**95.36**	92.05	94.70	94.74	88.88	90.58	88.88
kr_vs_kp	92.23	91.54	83.49	91.82	**93.56**	89.41	88.71
liver	**61.76**	**61.76**	57.65	10.29	60.90	4.12	57.07
mushroom	96.18	95.54	96.18	98.12	78.67	**99.37**	98.66
pima	72.11	71.18	68.95	20.79	**73.45**	63.94	69.26
promoters	**72.00**	**72.00**	/	/	29.13	42.50	63.00
spect	60.96	60.96	52.94	35.00	**64.42**	62.66	**64.42**
spectf	83.85	**85.00**	/	/	55.84	80.07	54.74
vote	**95.95**	**95.95**	90.00	93.33	95.40	95.40	95.40
wdbc	81.79	83.39	81.96	70.36	**95.79**	95.61	92.91
madelon	59.00	**60.85**	/	/	/	6.90	56.65
hiva	**93.70**	93.67	/	/	/	/	/
lymphoma	**77.27**	**77.27**	/	/	/	/	/
breast-cancer	**92.22**	91.48	/	/	/	/	/
ovarian-cancer	**92.86**	**92.86**	/	/	/	/	/
dorothea	**95.06**	**95.06**	/	/	/	/	/

TABLE IV. WIN/TIE/LOSS COUNTS

	MB-EP	**CAEP**	**SJEP**	**CBA**	**CMAR**	**CPAR**
CE-EP	6/11/3	17/2/1	19/0/1	16/0/4	17/0/3	17/0/3
MB-EP	/	16/1/3	17/0/3	15/0/5	17/0/3	17/0/3

C. *Comparison of the number of classification rules and running time of our two classifiers with the other four associative classifiers*

In this section, we compare the number of classification rules selected by CE-EP and MB-EP with the CAEP, CBA and CMAR classifiers, since those five associative classifiers all focus on generating classification rules with the support-confidence framework. Because of the page limit, Table V only reports the number of rules selected by five classifiers with the minimum support threshold up to 0.2 instead of all seven minimum supports used in the experiments and the best results are highlighted in bold.

TABLE V. COMPARISON OF THE NUMBER OF CLASSIFICATION RULES AMONG FIVE ASSOCIATIVE CLASSIFIERS (SUPPORT=0.2)

Datasets	**CE-EP**	**MB-EP**	**CAEP**	**CBA**	**CMAR**
australia	**5**	**5**	26	0	0
breast-w	**201**	**201**	**201**	365	832
heart	**14**	**14**	129	328	773
infant_mortality	**50**	58	/	109652	2646
ionosphere	**10**	43	10021	33023	3346
kr_vs_kp	**37**	146	614	818	151
liver	**4**	**4**	24	0	0
mushroom	**156**	353	388	3058	35025
pima	**18**	22	34	28	28
promoters	8	8	/	**5**	8
spect	**3**	**3**	43	125	890
spectf	**12**	58	/	98721	3236
vote	**18**	**18**	881	2610	11451
wdbc	**39**	57	2107	34100	33595
madelon	**7**	9	/	/	0
hiva	**14**	15	/	/	/
lymphoma	**33**	**33**	/	/	/
breast-cancer	**40**	116	/	/	/
ovarian_cancer	**17**	17	/	/	/
dorothea	**45**	60	/	/	/

As depicted in Table V, it is clear that the CE-EP and MB-EP classifiers select much fewer rules than the CAEP, CBA, and CMAR classifiers on all the datasets except for the *promoters* dataset. Those results further illustrate that extracting the classification rules from the space of direct causes and direct effects or the Markov blanket of the class label not only gets a much smaller set of rules but also achieves a higher predictive accuracy than the existing associative classifiers. Table V also indicates that even with very high-dimensional datasets, CE-EP and MB-EP only select a very small set of rules while CAEP, CBA, and CMAR cannot deal with those datasets, even with a rather high support value.

When the support value is up to 0.2, CBA and CMAR classifiers don't select any rules on the *australia* and *liver* datasets. Moreover, on the *madelon* dataset, CMAR also cannot select any rules. On the *promoters* dataset, although CBA and CMAR select a very small set of rules of 5 and 8 rules respectively, their corresponding classification accuracy is very low, only up to 15% and 2%.

Table VI compares the running time of our two classifiers with the CAEP, CBA, and CMAR classifiers with the minimum support thresholds up to 0.2. The best result is highlighted in bold face. The CE-EP and MB-EP classifiers are always faster than CAEP. Compared to CBA and CMAR, with the same minimum support threshold, on some small datasets, such as *breast-w, pima* and *vote*, CBA and CMAR are faster than CE-EP and MB-EP. On some datasets, such as *australia*, *live* and *promoters*, although CBA and CMAR are faster than CE-EP and MB-EP, CBA and CMAR almost don't select any rules and only achieve very low accuracy. The running time of CBA and CMAR fluctuates among different datasets while CE-EP has a very stable running time for both small and large datasets.

TABLE VI. RUNNING TIME OF FIVE ASSOCIATIVE CLASSIFIERS IN SECOND (SUPPORT=0.2)

Datasets	**CE-EP**	**MB-EP**	**CAEP**	**CBA**	**CMAR**
australia	12.58	12.58	14.67	**0.14**	3.072
breast-w	13.222	13.222	13.222	0.5	**0.187**
heart	10.28	10.28	13.08	**0.36**	0.438
infant_mortality	14.36	15.11	/	3557.41	**10.047**
ionosphere	**8.55**	11.56	19.5	208.67	51.516
kr_vs_kp	**13**	29.36	287.08	97.49	13.25
liver	**10**	**10**	10.51	0	0
mushroom	16.94	41.49	46.92	22.31	335.188
pima	10.49	10.49	10.49	0.25	**0.062**
promoters	12.39	13.39	/	**5.83**	5.9
spect	8.36	8.36	10.22	**0.86**	2.828
spectf	12.95	12.95	/	2375.63	**5.406**
vote	10.22	10.22	10.922	**1.36**	11.11
wdbc	**13.59**	13.84	48.39	111.42	226.219
madelon	**12.86**	13	/	/	20
hiva	**13.14**	13.25	/	/	/
lymphoma	**10.8**	10.8	/	/	/
breast-cancer	**46.7**	145	/	/	/
ovarian-cancer	**13.8**	20.8	/	/	/
dorothea	**160.4**	1789.2	/	/	/

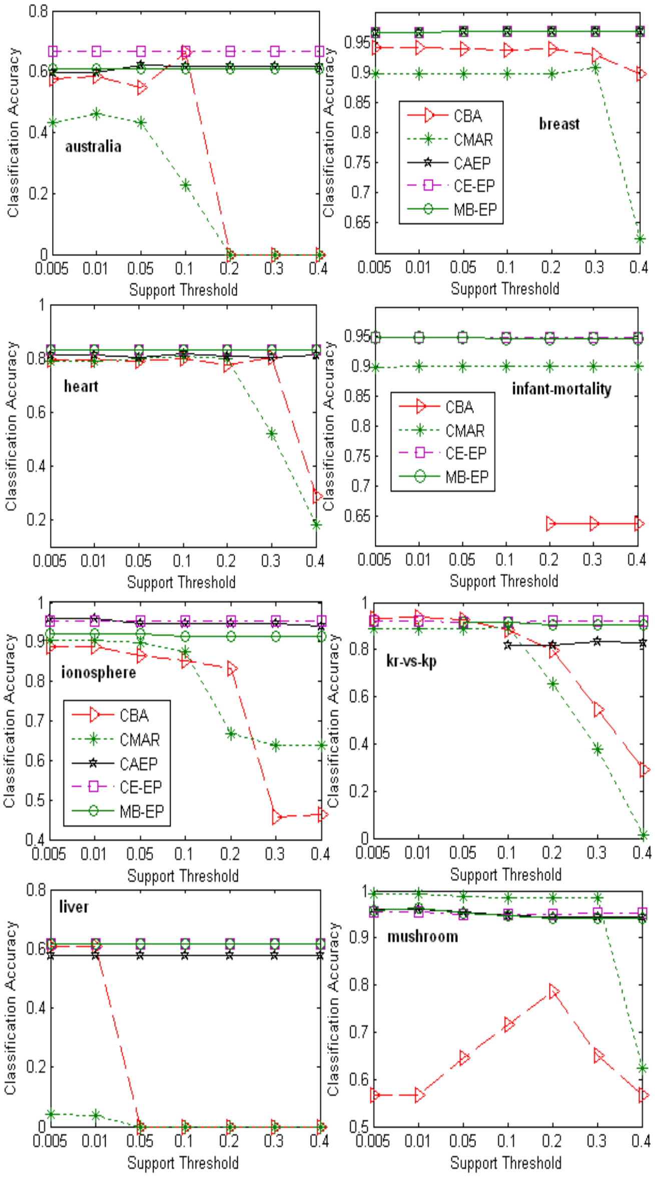
D. *Sensitivity Analysis on Support Thresholds*

To further explore the performance of CE-EP, MB-EP, CAEP, CBA and CMAR, we conduct a sensitivity analysis on the classification accuracy, number of selected rules, and running time of those five classifiers with seven minimum support threshold values.

(1) Sensitivity analysis on classification accuracy

We plot the classification accuracy of those five classifiers on the twenty datasets with seven support thresholds, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, and 0.4.

As illustrated in Fig.2, we can see that CE-EP, MB-EP and CAEP are less sensitive to the support thresholds than CBA and CMAR. 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 a higher accuracy under all the seven support thresholds than CAEP, CBA, and CMAR on all datasets except *mushroom* and *wdbc*.



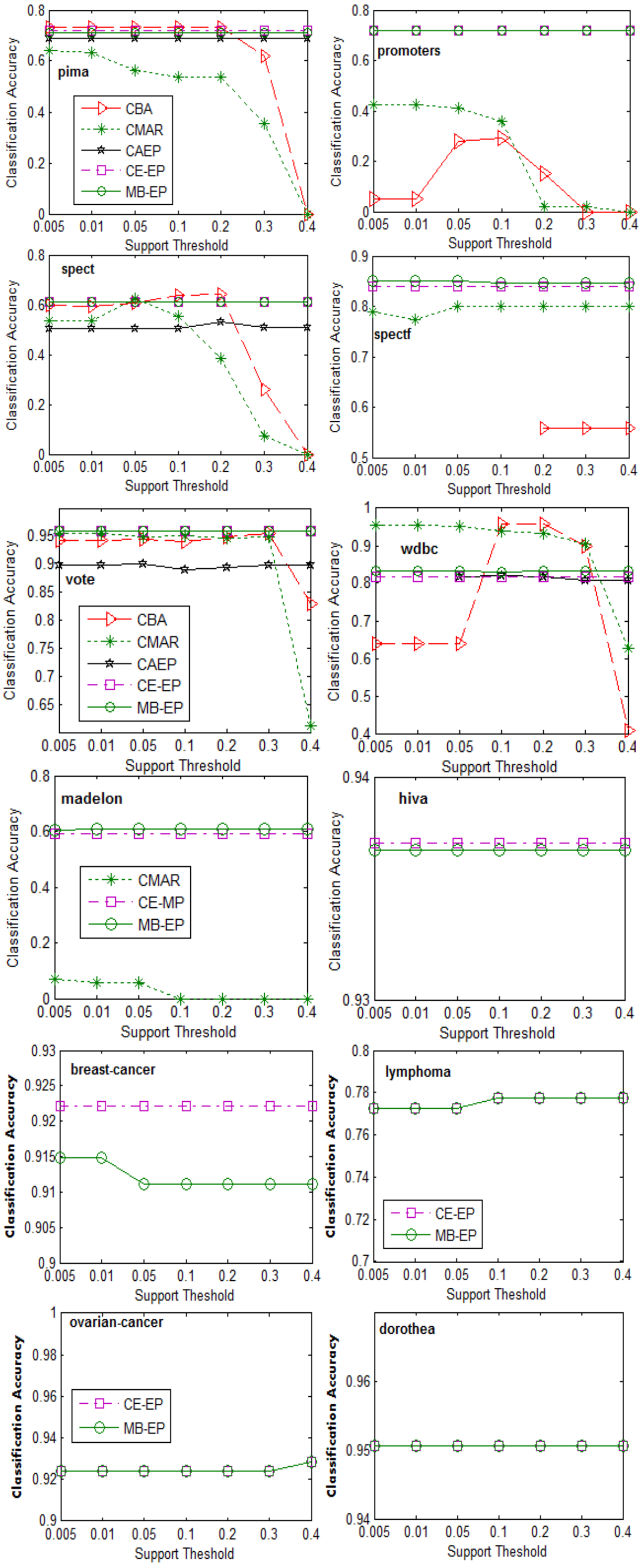


Figure 2. The classification accuracy of five associative classifers under different support threshold values on 20 datasets

On the other hand, CBA and CMAR are very sensitive to the support thresholds. For example, on all the datasets except for *infant-mortality* and *spectf*, when the support threshold is up to 0.3 or 0.4, the corresponding accuracy is greatly reduced. Moreover, on some datasets, the accuracy of CBA and CMAR is even reduced to 0, such as *australia, kr-vs-kp, liver, promoter* and *spect*. On the *infant-mortality* and *spectf* datasets, CBA can only work on the support value greater than 0.2.

From the above results, CE-EP and MB-EP are insensitive to the support thresholds. We have a further sensitivity analysis on the number of selected rules in the next section.

(2) *Sensitivity analysis on the number of selected rules*

We plot the number of selected rules of CE-EP, MB-EP, CAEP, CBA and CMAR on some datasets with the various support thresholds, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, and 0.4, in Fig. 3, instead of all the twenty datasets due to the page limit. From Fig.3, we can see that the numbers of selected rules of CBA and CMAR are very sensitive to the support thresholds. With a small support value, CBA and CMAR select a large number of classification rules. With the support value increasing, the number of selected rules is greatly reduced. Along with Fig. 2, we can conclude that when the support value is small, CBA and CMAR can obtain a large number of rules for classification to achieve good accuracy as shown in Fig. 2. When the support value is large, CBA and CMAR prune too many rules, including useful rules, and this results low classification accuracy. From Figures 2 and 3, it is clear that with a small support value, a large number of rules provide rich information for classification and make CBA and CMAR achieve high accuracy. Unfortunately, with a big support value, a large number of rules are generated, and they are difficult to store, retrieve and maintain for classification. For example, in Fig. 2, on the *infant-mortality* and *spectf* datasets, CBA doesn't work under a small support value because of a huge number of rules.

CAEP is less sensitive to the support value than CBA and CMAR. Therefore, CAEP also gets a more stable performance than CBA and CMAR under the different support values as shown in Fig. 2. But CAEP cannot deal with a high-dimensional dataset, even on the *spectf* dataset. As for CE-EP and MB-EP, especially CE-EP, under all seven support thresholds, CE-EP and MB-EP not only select a small set of rules, but also are insensitive to the support threshold, even on those very high-dimensional datasets. Therefore, those results further demonstrate that the importance of introducing causality into associative classification. More especially, when CE-EP and MB-EP mine the classification rules from the space of the direct causes and direct effects or Markov blanket of the class label, they can achieve strongly predictive rules no matter whether the support value is small or large. This also explains why the accuracy of CE-EP and MB-EP remains stable under the different support thresholds as shown in Fig. 2.

Although Fig. 3 only contains eight datasets, the results on the remaining twelve datasets have a similar trend. Therefore, in a word, among those five associative classifiers, as for the number of selected rules, CE-EP and MB-EP are the most insensitive classifiers to the support thresholds while CBA and CMAR are the most sensitive ones. CAEP is less sensitive than CBA and CMAR.

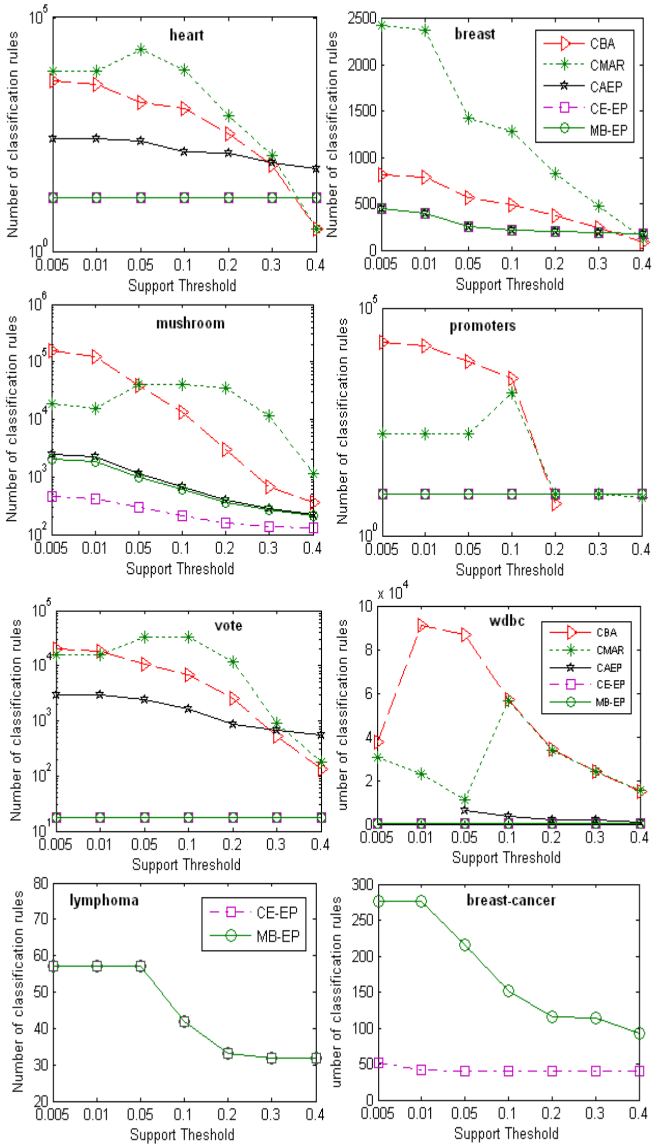


Figure 3. Number of selected rules with varied support thresholds

(3) Sensitivity analysis on running time

Fig. 4 shows the running time of the five classifiers under seven support values, on the same eight datasets as in Fig.3. From Fig. 4, we can see that the running time of CAEP, CBA, and CMAR is very sensitive to different support thresholds, especially CBA and CMAR. In Table VI, we report the running time of those five classifiers under the support value up to 0.2. CBA and CMAR are much faster than CE-EP and MB-EP on some small datasets. The running time of CE-EP and MB-EP is also insensitive to the support values while the running time of the CBA and CMAR classifiers fluctuates under the different support values.

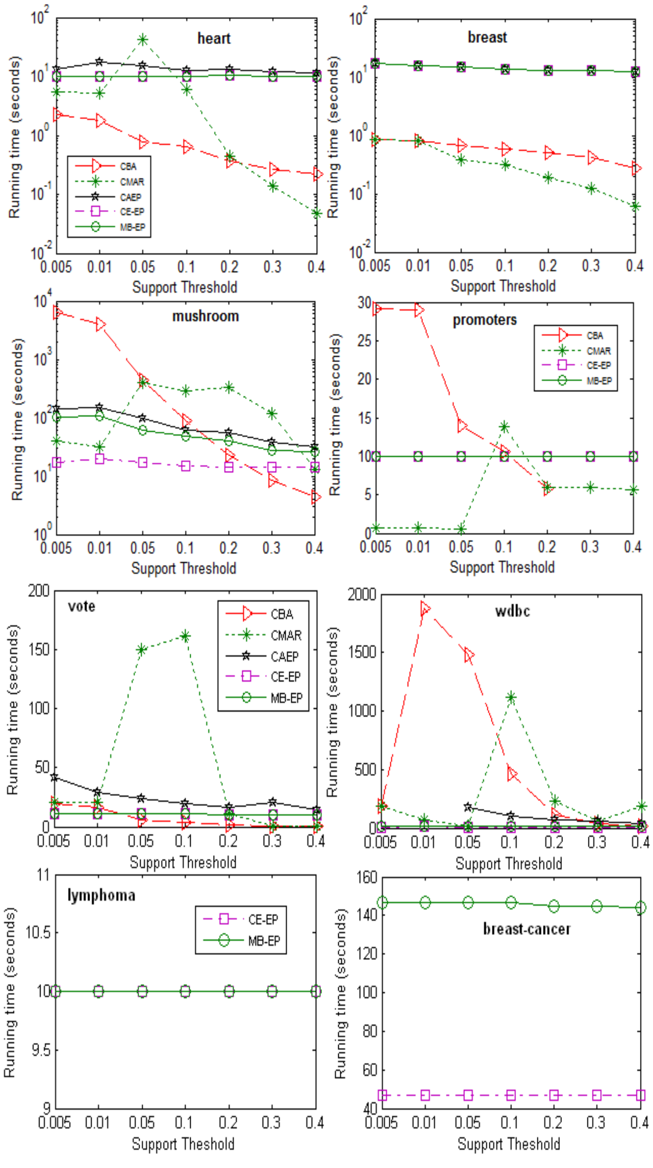


Figure 4. Running time with varied support thresholds

E. Comparison of classification accuracy of our classifiers with the other state-of-the-art classifiers

Table VII compares the accuracy of our two classifiers against well-known classifiers such as decision tree J48, Naive Bayes (NB), ensemble classifiers Bagging and Boosting implemented in Weka. We also present a win/tie/loss summary in Table VIII to compare the overall performance of our two classifiers against each other. From Tables VII and VIII, both CE-EP and MB-EP achieve the highest accuracy on the 9 data sets of the 20 data sets, respectively. On the remaining data sets, while our two classifiers are not the best, they achieve the accuracy very close to the highest on most of these 11 data sets. Our two classifiers are more accurate than NB and J48 in general. Moreover, they are also slightly better than the more advanced ensemble classifiers: Bagging and Boosting.

TABLE VII. A COMPARISON OF PREDICTIVE ACCURACY OF OUR CLASSIFIERS WITH THE OTHER STATE-OF-THE-ART CLASSIFIERS

Datasets	**CE-EP**	**MB-EP**	**J48**	**NB**	**Bagging**	**Boosting**
australia	66.47	60.88	66.23	66.23	65.07	**67.10**
breast-w	**96.88**	**96.88**	94.56	95.99	95.57	94.85
heart	**83.33**	**83.33**	78.52	81.85	79.26	81.485
infant_mortality	94.88	94.78	95.39	91.91	**95.65**	95.43
ionosphere	**95.36**	92.05	90.73	93.38	90.73	91.39
kr_vs_kp	92.23	91.54	**99.31**	83.92	99.22	93.84
liver	**61.76**	**61.76**	60.00	61.16	59.71	60.87
mushroom	96.18	95.54	**100.00**	85.68	**100.00**	98.44
pima	72.11	71.18	72.00	70.44	72.40	**73.18**
promoters	72.00	72.00	63.21	**74.53**	60.38	66.04
spect	**60.96**	**60.96**	49.20	49.73	55.085	59.89
spectf	83.85	85.00	62.57	86.63	**90.375**	75.40
vote	**95.95**	**95.95**	94.01	94.23	95.40	82.87
wdbc	81.79	**83.39**	75.92	77.68	80.32	76.27
madelon	59.00	60.85	57.50	59.20	**62.20**	60.50
hiva	93.7	93.67	96.39	87.06	/	**96.47**
lymphoma	**77.27**	**77.27**	70.93	68.28	64.76	60.79
breast-cancer	92.22	91.48	80.77	**93.01**	84.97	83.57
ovarian-cancer	**92.86**	**92.86**	89.35	70.83	88.89	90.74
dorothea	**95.06**	**95.06**	/	/	/	/

TABLE VIII. WIN/TIE/LOSS COUNTS

	NB	**J48**	**Bagging**	**Boosting**
CE-EP	16/1/3	15/1/4	14/0/6	13/0/7
MB-EP	15/0/5	14/0/6	13/0/7	14/0/6

V. CONCLUSIONS

How to deal with a huge number of association rules from a high-dimensional data set is a challenging issue in associative classification. Meanwhile, how to deal with high sensitivity to the minimal support threshold is another challenging problem. In this paper, we bring causality into play when designing associative classifiers and propose a new framework for associative classification. In order to validate our framework, we use EP-based classifiers as baseline methods to instantiate our framework and propose two causal associative classifiers: CE-EP and MB-EP. Experimental results have demonstrated the effectiveness and efficiency of our proposed framework.

Meanwhile, we have only instantiated our framework on EP-based classifiers. Exploring our framework in general association rule classifiers is our future work.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (60975034, 61070131, 61175051 and 61005007), the US National Science Foundation (CCF-0905337), and the US NASA Research Award (NNX09AK86G).

REFERENCES

[1] H. Alhammady and K. Ramamohanarao. Using Emerging Patterns and Decision Trees in Rare-Class Classification. ICDM '04: 315-318.

[2] C. F. Aliferis, I. Tsamardinos, A. Statnikov and L.E. Brown. Causal Explorer: a Causal Probabilistic Network Learning Toolkit for Biomedical Discovery. METMBS'03, 2003.

[3] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani and X.D. Koutsoukos. 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(Jan):171−234, 2010.

[4] E. Baralis, S. Chiusano, and P. Garza. A Lazy Approach to Associative Classification. IEEE Trans. Knowl. Data Eng. 20(2): 156-171, 2008.

[5] N.M Blachman. The Amount of Information That y Gives about X. IEEE transactions on Information Theory. 14(1), 27-31,1968.

[6] G. Bontempi and P. E. Meyer. Causal Filter Selection in Microarray Data. ICML'10, 2010.

[7] L. Chen and G. Dong. Masquerader Detection Using Oclep: One-class Classification Using Length Statistics of Emerging Patterns. WAIMW'06, 2006.

[8] G. F. Cooper and E. Herskovits. A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning, 9: 309-347 (1992)

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

[10] G. Dong and J. Li. Efficient Mining of Emerging Patterns: Discovering Trends and Differences. KDD'99, 1999.

[11] G. Dong, X. Zhang, L. Wong and J. Li. CAEP: Classification by Aggregating Emerging Patterns. DS'99, 30–42, 1999.

[12] H. Fan and K. Ramamohanarao. 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, 2006.

[13] I. Guyon, C. Aliferis and A. Elisseeff. Computational Methods of Feature Selection, in chapter of Causal Feature Selection. pp. 63−86. Chapman and Hall, 2007.

[14] J. Li, G. Dong and K. Ramamohanarao. Making Use of the Most Expressive Jumping Emerging Patterns for Classification. Knowledge and Information Systems, 3 (2001), 131−145.

[15] W. Li, J. Han and J. Pei. CMAR: Accurate and Efficient Classification Based on Multiple-class Association Rule. ICDM'01, pp. 369–376, 2001.

[16] B. Liu,W. Hsu and Y. Ma. Integrating Classification and Association Rule Mining. KDD'98, 80-86, 1998.

[17] T.S Ngo, M. Feng, G. Liu and L. Wong. Efficiently Finding the Best Parameter for the Emerging Pattern-Based Classifier PCL. PAKDD'10, 2010.

[18] J. Pearl. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Francisco, California, 2nd edition, 1991.

[19] K. Ramamohanarao and H. Fan. Patterns Based Classifiers. World Wide Web 10(1), 71−83 (2007).

[20] C. Silverstein, S. Brin and R. Motwani. Scalable Techniques for Mining Causal Structures. In proceedings of the 24th VLDB Conference, New York, USA, 1998.

[21] P. Spirtes, C. Glymour and R. Scheines. Causation, Prediction, and Search (Second ed.). The MIT press, 2000.

[22] P. Smyth and R.M. Goodman. Rule Induction Using Information Theory. In: Piatetsky-Shapiro, G. and Frawley, W.J. (eds.), Knowledge Discovery in Databases. AAAI Press, pp. 159-176(1991).

[23] I. Tsamardinos, L. E. Brown and C. F. Aliferis. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. Machine Learning, 651:31-78, 2006.

[24] THE LUCS-KDD SOFTWARE. http://www.csc.liv.ac.uk/~frans/KDD/Software.

[25] X. Yin and J. Han. CPAR: Classification Based on Predictive Association Rule. SDM'03, 369–376, 2003.