2013 IEEE 13th International Conference on Data Mining

Markov Blanket Feature Selection with Non-Faithful Data Distributions

Kui Yu[1], Xindong Wu[1,2], Zan Zhang[1], Yang Mu[3], Hao Wang[1], and Wei Ding[3]

[1]Department of Computer Science
Hefei University of Technology
Hefei, 23009, 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; zhangzan99@163.com; jsjxwangh@hfut.edu.cn; {ding,yangmu}@cs.umb.edu

***Abstract*—In faithful Bayesian networks, the Markov blanket of the class attribute is a unique and minimal feature subset for optimal feature selection. However, little attention has been paid to Markov blanket feature selection in a non-faithful environment which widely exists in the real world. To tackle this issue, in this paper, we deal with non-faithful data distributions and propose the concept of representative sets instead of Markov blankets. With a standard sparse group lasso for selection of features from the representative sets, we design an effective algorithm, SRS, for Markov blanket feature Selection via Representative Sets with non-faithful data distributions. Empirical studies demonstrate that SRS outperforms the state-of-the-art Markov blanket feature selectors and other well-established feature selection methods.**

Keywords-Feature selection; Markov blankets; Faithful Bayesian networks; Representative sets; Sparse group lasso

I. Introduction

Markov blankets in Bayesian networks were first introduced by Pearl [13], and in faithful Bayesian networks, for every node X, its Markov blanket is the set of parents, children and spouses (parents of the children of X) as shown in Figure 1 [6]. Koller and Sahami [7] first introduced Markov blankets for feature selection defined by feature relevance. In feature relevance with respect to the class attribute, an input feature can be classified into a strongly relevant, irrelevant, redundant, or non-redundant feature, and a Markov blanket should include strongly relevant and non-redundant features [22]. However, using feature relevance to exactly determine Markov blankets is very difficult because of a limited sample size and noise in the data [7].

To tackle this issue, Tsamardinos and Aliferis [19] provided theoretical results that link the concepts of feature relevance in feature selection and Markov blankets in Bayesian networks. Their theoretical results proved that if a probability distribution can be faithfully represented by a Bayesian network, then the Markov blanket of the class attribute in the Bayesian network is not only unique but also the solution to feature selection. With those theoretical results, Markov blanket feature selection has attracted much attention in recent years [1, 14, 16, 24]. Tsamardinos and Aliferis [19] proposed IAMB that returns the Markov blanket of any target node in a faithful Bayesian network without learning a complete Bayesian network, even with hundreds of thousands of features. However, it requires a sample size exponential in the size of a Markov blanket. More recent variations of IAMB include PCMB (Parent-Children Markov Blanket) [14], MMMB (Max-Min Markov Blanket) [1, 20], and HITON-MB [1], proposed to conquer the data inefficiency problem of IAMB. Moreover, Aliferis et al. [1] demonstrated that Markov blanket feature selection outperforms most of the state-of-the-art feature selection algorithms.

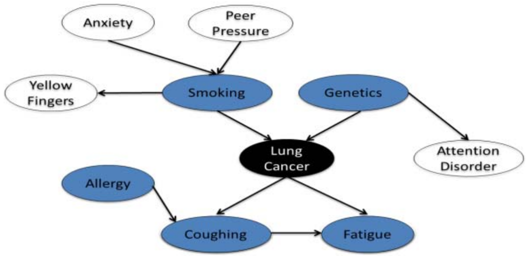

Figure 1. The Markov blanket (in blue) of Lung Cancer

Meanwhile, the previous studies mentioned above typically assume that a data distribution and an underlying Bayesian network which models that domain are faithful to each other. This assumption relies on an important theoretical result that if a joint probability distribution $\mathbb{P}$ satisfies the intersection property (see Section III), then it is guaranteed to have a unique Markov blanket of the target feature/variable [13]. Moreover, the probability distribution $\mathbb{P}$ that is faithful to an underlying Bayesian network also satisfies the intersection property [13-14]. Thus, in a faithful Bayesian network, it is guaranteed to have a unique Markov blanket of any target node [12, 19].

However, in some real-life distributions, Markov blankets of a target feature are not unique and may vary in size due to various factors, such as (but not limited to) small sample size, noise in data, hidden variables, and data pre-processing [14, 17]. This makes real data in many cases violate the intersection property or faithfulness condition. Thus, it is strict yet difficult to handle data in faithfulness conditions, and dealing with real data with non-faithful distributions is evidently more meaningful.

For example, Figure 2 gives an example to illustrate the multiple Markov blanket problem in real data of the arcene (cancer benchmark) dataset (with 100 instances and 10000 features) from the NIPS 2003 feature selection challenge. In Figure 2, IAMB, HITON-MB, PCMB and MMMB can only discover a single Markov blanket. Different from those four algorithms, KIAMB [14] can find multiple Markov blankets, by employing a stochastic

1550-4786/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDM.2013.154

search heuristic that repeatedly disrupts the order in which features are selected for inclusion into a Markov blanket with the probability p, thereby introducing a chance of identifying alternative Markov blankets of a target feature.

We run KIAMB 100 times to attain 100 Markov blankets (the parameter p is set 0.6), respectively. By using Decision Tree J48 and Knn classifiers, Figure 2 gives a summary of the prediction accuracies of those 100 Markov blankets and those of the Markov blankets selected by IAMB, HITON-MB, PCMB and MMMB.

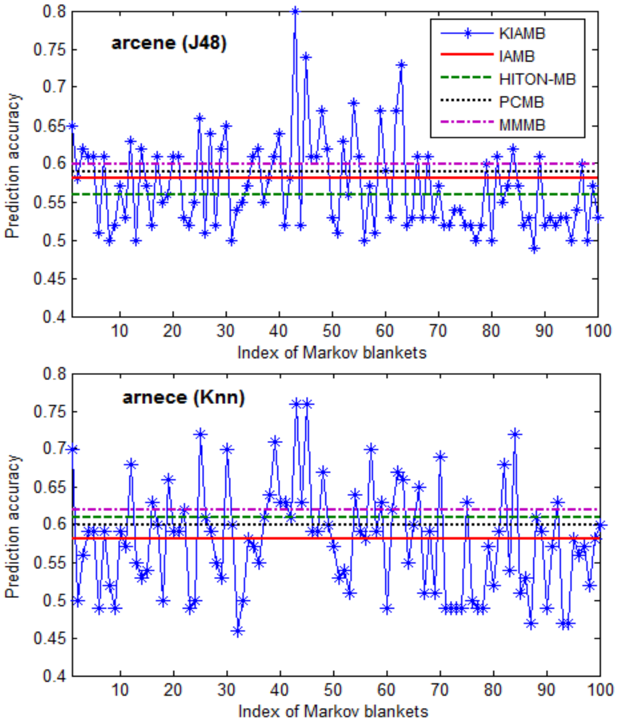

Figure 2. Prediction accuracy on the arcene dataset

From Figure 2, we can see that the Markov blankets identified by the existing Markov blanket algorithms are not the optimal feature subsets for feature selection, compared to the Markov blankets discovered by KIAMB.

Therefore, Markov blanket feature selection in non-faithful data distributions needs further attention. However, when we deal with non-faithful data distributions, the challenges are three-fold as follows.

Firstly, if the real data distributions contain multiple Markov blankets, we don't know the exact number of Markov blankets of any target feature in real data.

Secondly, even if we know all Markov blankets of a target feature, it is very expensive, or even impossible to discover all of them since the number of Markov blankets can grow exponentially in the number of features in the underlying Bayesian network.

And thirdly, with multiple Markov blankets, the Markov blankets discovered by the existing Markov blanket feature selection algorithms might not be the optimal solutions to feature selection, and thus, how can we efficiently find a best Markov blanket as the solution to feature selection from a large and even exponential number of Markov blankets in a non-faithful distribution?

To address those problems, the main contributions of the paper are as follows.

(1) We propose the concept of a representative set instead of a Markov blanket to focus on a feature space of all possible Markov blankets, instead of an exhaustive search over an unknown and even exponential number of Markov blankets in real data.

(2) To define representative sets, we extend the theoretical results provided by Tsamardinos and Aliferis [19] about Markov blankets and feature relevance. Our theoretical analysis and empirical study both show that in Markov blankets only parents and children correspond to strongly relevant features in feature relevance. With this theoretical result, a representative set is defined as parents or children in a Markov blanket and their corresponding correlated features.

(3) With representative sets, we employ the standard sparse group lasso approach, and design an effective algorithm, SRS, for Markov blanket feature ***S***election via ***R***epresentative ***S***ets to process data in non-faithful distributions. Empirical results on high-dimensional datasets show that the SRS algorithm outperforms state-of-the-art Markov blanket feature selectors and other well-established feature selection methods.

II. RELATED WORK

Feature selection aims to reduce the computational complexity without performance degradation by removing irrelevant and redundant features. The major effort is to maximize relevance and minimize redundancy among the selected features for classification. For instance, the mRMR (Minimum Redundancy Maximum Relevance) algorithm proposed by [15] while the FCBF (Fast Correlation Based Filter) algorithm proposed by [22]. Recently, Cheng et al. [3] presented a Fisher-Markov filter method to identify a maximally separable feature subset using the Fisher's discriminant analysis and the Markov random fields (MRFs). Brown et al. [2] proposed a unifying framework to bring almost two decades of research on heuristic scoring criteria through a novel interpretation of information theoretic feature selection as an optimization of the conditional likelihood. Zhao et al. [26] proposed a framework to unify different criteria for handling feature redundancies.

Markov blanket feature selection as an emerging successful class of filter methods presents a solution of the feature selection problem by discovery of a Markov blanket of the class attribute [1], and thus it has attracted much attention [14, 23].

Margaritis and Thrun invented the first yet sound Markov blanket discovery algorithm, the GS algorithm with the intent to discover the Markov blanket for the purpose of speeding up global Bayesian network learning [11]. But the GS algorithm requires the number of instances exponentially to the size of the Markov blanket, and this makes it impractical for many real datasets.

To conquer this drawback of the GS algorithm and apply the concept of Markov blanket to the feature selection task, Tsamardinos and Aliferis [19] provided theoretical results that link feature relevance as defined by Kohavi and John [8] and the Markov blanket in faithful Bayesian networks. And then, they proposed a modified version of the GS algorithm, called the IAMB algorithm for feature selection, which guarantees to find the actual Markov blanket given enough training data and is more sample efficient than GS [19]. However, the IAMB algorithm still requires a sample size exponential in the size of a Markov blanket. Thus, HITON-MB and MMMB were introduced without requiring samples exponential in the size of the Markov blanket. Different from GS and IAMB, HITON-MB and MMMB take two steps to find the Markov blanket of a target node: (1) discovering the parents and children of the target node; and then (2) identifying its spouses based on Step 1. As an efficient implementation of Step 1, two major algorithms HITON-PC and MMPC were introduced [1, 20]. Following the idea of MMMB, PCMB was also proposed to conquer the data inefficiency problem of IAMB [14].

However, the algorithms mentioned above are well-established only for selection of a single Markov blanket problem by handling data in faithful data distributions, and little research has been done in the development of algorithms for dealing with Markov blanket feature selection problem with non-faithful data distributions.

A naïve approach for handling Markov blanket feature selection in non-faithful data distributions involves first clustering all features into multiple clusters, and then randomly sampling a representative from each cluster. But this strategy is intractable since the computation is intensive for high feature dimensions, and features in each cluster don't indicate they are correlated in terms of feature relevance [21]. Peña et al. [14] proposed a stochastic Markov blanket algorithm, called KIAMB, that involves running multiple times initialized with a random seed. Recently, among the most notable advances in the field is that Statnikov et al. [17] proposed the TIE* (Target Information Equivalence) algorithm that can discover all Markov blankets in a non-faithful data distribution. But TIE* is very expensive or prohibitive when the number of Markov blankets grows exponentially in the number of features in the network.

III. Notations and Definitions

In the following sections, let $F = \{F_1, F_2, \ldots, F_n\}$ represent a full set of features, C denote the class attribute, and $F - \{F_i\}$ represent the feature subset excluding F_i.

A. Bayesian networks

Definition 1 (Conditional Independence) Two distinct features $F_i \in F$ and $F_k \in F$ are conditionally independent on a subset $S \subseteq F - \{F_i \cup F_k\}$($\mathrm{Ind}(F_i, F_j|S)$ for short), iff $P(F_i|F_k, S) = P(F_i|S)$ or $P(F_k|F_i, S) = P(F_k|S)$.

Definition 2 (Bayesian Networks) [13] Let $\mathbb{P}$ be a discrete joint probability distribution of a set of random nodes (features) F via a directed acyclic graph $\mathbb{G}$. We call the triplet $< F, \mathbb{G}, \mathbb{P} >$ a (discrete) Bayesian network if $< F, \mathbb{G}, \mathbb{P} >$ satisfies the **Markov condition**: every node is independent of any subset of its non-descendant nodes conditioned on its parents. (A simple Bayesian network of Lung Cancer has been shown in Figure 1.)

Theorem 1 [13] Let X, Y, Z, and W be any four subsets of features from F and a joint probability distribution $\mathbb{P}$ is strictly positive. Then the following intersection property holds in $\mathbb{P}$ over the feature set F:

$$\mathrm{Ind}(X, Y|Z \cup W) \text{ and } \mathrm{Ind}(X, W|Z \cup Y) \Rightarrow \mathrm{Ind}(X, (Y \cup W)|Z).$$

Theorem 2 [13] If a joint probability distribution $\mathbb{P}$ over the feature set F satisfies the intersection property, then for each $X \in F$, there exists a unique Markov blanket of X.

Definition 3 (Faithfulness) [13] A Bayesian network satisfies the faithfulness condition if and only if every conditional independence entailed by the directed acyclic graph $\mathbb{G}$ is also present in the joint probability $\mathbb{P}$.

Theorem 3 [13] If $\mathbb{P}$ is faithful to $\mathbb{G}$, then $\mathbb{P}$ satisfies the intersection property.

With Theorems 2 and 3, the concept of Markov blankets in faithful Bayesian networks is defined as follows.

Definition 4 (MB: Markov Blanket) [19] In faithful Bayesian networks, for every node F_i, its Markov blanket is unique with the set of parents, children and spouses of F_i.

Definition 5 (Collider) [13] A node $F_i \in F$ of a path p is a collider if p contains two incoming edges into F_i.

Proposition 1 [13] A path p from node $F_i \in F$ to node $F_k \in F$ is blocked by a set of nodes $S \subset F$, if there is a node $F_m \in F$ on p for which one of the following two conditions hold: (a) F_m is not a collider and $F_m \in S$, or (b) F_m is a collider and neither F_m or its descendants are in S.

Definition 6 (D-separation) [13] Two nodes $F_i \in F$ and $F_k \in F$ are d-separated by $S \subset F$ in graph G if and only if every path from F_i to F_k is blocked by S.

Theorem 4 [19] In faithful Bayesian networks, d-separation captures all conditional dependence and independence relations that are encoded in the graph which implies that two nodes are d-separated with each other given a subset S, iff they are conditionally independent conditioned on S.

B. Feature relevance in Feature Selection

In this section, we introduce the concepts of feature relevance proposed by Kohavi and John [8].

Definition 7 (Strong Relevance) A feature F_i is strongly relevant to C iff

$$\forall S \subseteq F - \{F_i\} \text{ s.t. } P(C|S) \neq P(C|S, F_i).$$

Definition 8 (Weak Relevance) A feature F_i is weakly relevant to C iff it is not strongly relevant, and

$$\exists S \subset F - \{F_i\} \text{ s.t. } P(C|S) \neq P(C|S, F_i)$$

Definition 9 (Irrelevance) A feature F_i is irrelevant to C iff it is neither strongly nor weakly relevant, and

$$\forall S \subseteq F - \{F_i\} \text{ s.t. } P(C|S, F_i) = P(C|S)$$

Yu and Liu [22] divided weakly relevant features into redundant features and non-redundant features.

Definition 10 (Redundant Features) Assuming $S \subseteq F - \{F_i\}$ as the current feature set, a feature F_i is redundant and hence should be discarded from S, iff it is weakly relevant and has a MB within S.

Accordingly, an optimal feature subset should consist of strongly relevant features and non-redundant features.

Tsamardinos and Aliferis [19] proved the following theorem to link feature relevance in feature selection and the Markov blanket in faithful Bayesian networks.

Theorem 5 A feature $X \in F$ is strongly relevant, iff it belongs to the Markov blanket of the class attribute in a faithful Bayesian network.

Theorem 5 confirms that the Markov blanket of the class attribute in faithful Bayesian networks is not only unique but also the solution to feature selection.

IV. Selection of Features via Representative Sets

A. Representative Sets

As stated above, when a data set doesn't satisfy the intersection property or faithful distribution, it may have multiple Markov blankets of a target feature. Since the number of Markov blankets can grow exponentially in the number of features in the underlying Bayesian network, the discovery of all Markov blankets in non-faithful data distribution and picking up a best Markov blanket for feature selection are very expensive, and sometimes are infeasible with high feature dimensions. This motivated us to propose a novel algorithm to solve the problem of multiple Markov blanket selection in real data.

In feature selection, redundant features can replace others in a feature subset, and this is why a MB need not be unique in real data. Figure 2 illustrates that redundant features discarded by a Markov blanket feature selection algorithm actually carries a stronger predictive ability than the selected features in Markov blankets. Feature redundancy usually is defined by means of feature correlation [9], thus we call those redundant yet discarded features as correlated features with respect to the selected features in Markov blankets.

With correlated features, the feature space of all possible Markov blankets may consist of features in a Markov blanket and their corresponding correlated features. It is an efficient way to discover the feature space of all possible Markov blankets instead of an exhaustive search over an unknown yet even exponential number of Markov blankets in real data. With those observations, by dealing with data in non-faithful distributions, we extend the concept of Markov blankets, and propose the concept of representative sets for defining the feature space of all possible Markov blankets.

Definition 11 (Representative Sets) A representative set consists of a feature in a Markov blanket and its corresponding correlated features.

Different from Markov blankets, each member in representative sets isn't a single feature any longer, but a feature subset. With Definition 11, now the problem is how can we obtain representative sets in an efficient way?

To tackle this problem, we further extend the theoretical result in Theorem 5 which illustrates that the features in a MB are all strongly relevant features.

Lemma 1 [19] In Bayesian networks, with two nodes $F_i \in F$ and $F_k \in F$, if F_i and F_k are never d-separated given any subset of the nodes within $S \subseteq F - \{F_i \cup F_k\}$, iff there must exist a direct edge between F_i and F_k.

Lemma 2 [19] In Bayesian networks, with two nodes $F_k \in F$ and $F_i \in F$, and their common child $F_m \in F$, if F_i has no direct edge to F_k, F_i and F_k cannot be d-separated given any subset of the nodes within $S \subseteq F - \{F_k \cup F_i\}$ that contains F_m.

Proposition 2 A feature $F_i \in F$ is a strongly relevant feature, iff F_i belongs to the set of parents and children of the class attribute C in a faithful Bayesian network.

Proof: Assume F_i is a strongly relevant feature. By Definition 7, F_i and C are conditionally dependent given any subset S within $F - \{F_i\}$, that is, $\forall S \subseteq F - \{F_i\} \text{ s.t. } P(C|S) \neq P(C|S, F_i)$. From Theorem 4, this implies that F_i and C are never d-separated by any subset within F excluding F_i. By Lemma 1, we conclude that in Bayesian networks, feature F_i belongs to the set of parents and children of C.

Conversely, in faithful Bayesian networks, if node F_i belongs to the set of parents and children of C, then F_i and C are never d-separated by any subset within F excluding F_i by Lemma 1. Accordingly, by Theorem 4, we come to a conclusion that node F_i coincides with the definition of a strongly relevant feature in Section III.B. □

Proposition 3 In faithful Bayesian networks, if node $F_i \in F$ belongs to spouses of the class attribute C and F_i does not have a direct edge to C, then F_i is a non-redundant feature.

Proof: Since F_i is a spouse of C, and F_i has not a direct edge to C, in the path p from F_i to C, the common child of both C and F_i, named F_m, is a collider. Thus, this implies that F_i and C are d-separated by a subset that doesn't contain F_m by Lemma 2. Then we can obtain the term: $\exists S \subset F - \{F_m \cup F_i\}, P(C| F_i, S) = P(C|S)$ by Theorem 4. Thus, F_i is not a strongly relevant feature. On the other hand, according to Lemma 2, we can conclude that F_i and C cannot be d-separated by any subset that contains F_m, that is, the following term holds by Theorem 4:

$$Z = F - \{F_m \cup F_i\}, \forall S \subset Z \cup F_m, P(C, |S) \neq P(C|S, F_i).$$

Thus, F_i is weakly relevant to C, and we cannot find a Markov blanket in Z that contains F_m to make F_i redundant to C, hence F_i should be a non-redundant feature with respect to C. □

With Propositions 2 and 3, we extend Theorem 5 and obtain Theorem 6.

Theorem 6 In faithful Bayesian networks, the Markov blanket of the class attribute includes (1) parents and children corresponding to strongly relevant features; and (2) spouses corresponding to non-redundant features.

With Theorem 6, the next step is to determine the correlated features related to the features in an MB.

It is clear that node X and its parents (or children) are conditionally dependent given any subset within the remaining nodes in a Bayesian network. Thus, node X and its parents (or children) are correlated to each other. In addition, by the Markov condition, X is independent of any subset of its non-descendant nodes conditioned on its parents, but not its children nodes, so the features directly correlated to X can be defined in Definition 12.

Definition 12 The features directly correlated to X are defined as the set of parents and children of node X in a Bayesian network.

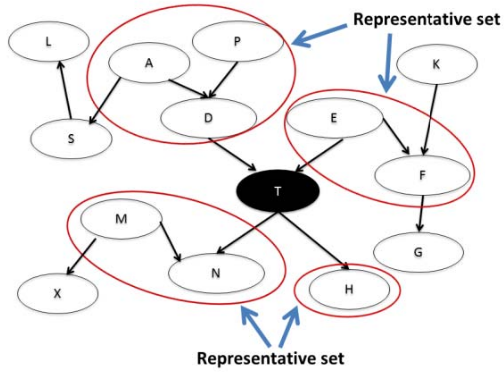

Figure 3. Representative sets related to node "T"

Proposition 4 A representative set can be obtained from Bayesian networks, including a parent or a child of the class attribute (a strongly relevant feature), and the set of parents and children of this parent or child (correlated features).

But why don't we treat a spouse in MBs and its parents and children as a representative set? The reason is that a spouse as a non-redundant feature has already been selected in a representative set. For instance, Figure 3 shows four representative sets related to node "T" in red color, considering node "T" as the class attribute.

B. *Selection of Features from Representative Sets*

With representative sets, the problem becomes how to select a best subset from representative sets? To handle this problem, we need to solve how to simultaneously optimize selections within each representative set as well as between those sets to achieve a feature subset that maximizes the predictive power to the class attribute. Suppose the predictive power of a feature subset to the class attribute is measured by a loss function $\mathcal{L}(\cdot)$, and $G = \{G_1, G_2, \ldots, G_K\}$ represents K representative sets with respect to K strongly relevant features. Each element $G_i (i = 1 \ldots K)$ in G incorporates the ith strongly relevant feature and its corresponding correlated features. Then our solution has two following steps.

The first step is to identify which representative sets have the most predictive power to the class attribute, and we formulate this step as follows.

$$\beta^* = \arg\min_\beta \mathcal{L}(\beta, G, C) + \lambda_1 \sum_{i=1}^{K} \Omega_1(\beta_i) \qquad (1)$$

where $\beta = \{\beta_1, \beta_2, \ldots, \beta_K\}$ is the coefficient vector for all representative sets, and β_i is the coefficient vector corresponding to G_i, C is the class attribute vector, $\Omega_1(\beta_i)$ is the regularization term to control the complexity of β_i, the parameter λ_1 controls the selection of set, and if β_i=0, then the ith set will be excluded entirely.

The second step is to calculate one or a few feature(s) to be chosen from each selected representative set, which is expected to contribute the most predictive power to the class attribute. The objective function in Eq. (1) is then further formulated as,

$$\beta^* = \arg\min_\beta \mathcal{L}(\beta, G, C) + \lambda_1 \sum_{i=1}^{K} \Omega_1(\beta_i) + \lambda_2 \Omega_2(\beta) \ (2)$$

where $\Omega_2(\beta)$ penalizes the complexity of β.The parameter λ_2 adjusts the individual feature coefficient β to select features within each set and if there is a coefficient in β up to 0, then the corresponding feature is discarded.

To solve Eq.(2), we adopt the sparse group lasso approach which is an extension of Lasso and group Lasso [18, 25]. The sparse group lasso can yield a best feature subset within each group and between groups simultaneously [4], and in our case representative sets are groups. The sparse group lasso penalizes the coefficients, β_i and β, by adding two constraints: 1) an ℓ_2 norm for $\Omega_1(\beta_i)$ to constrain the coefficients between sets, and 2) an ℓ_1 norm for $\Omega_2(\beta)$ to penalize feature coefficients within a set. Finally, if we employ the least square loss as the loss function $\mathcal{L}(\cdot)$, Eq.(2) can be rewritten as the following objective function:

$$\min_\beta \left\| C - \sum_{i=1}^{K} G_i \beta_i \right\|_2^2 + \lambda_1 \sum_{i=1}^{K} ||\beta_i||_2 + \lambda_2 ||\beta||_1 \ \ (3)$$

Thus we cast our problem as the standard sparse group lasso model. Note that Eq.(3) can be solved using any standard sparse group lasso algorithm.

C. *The Proposed Algorithm*

We design SRS (**S**election via **R**epresentative **S**ets) to implement the ideas discussed in Sections IV.A and IV.B. **SF** denotes a set of strongly relevant features, **G** denotes representative sets, and **PC(T)** denotes the parents and children of a target feature T in Figure 4.

The SRS Algorithm

```
Input: data with the class attribute C, $\lambda_1$ and $\lambda_2$
Step 1: Identify representative sets G
  (1) SF=Get-PC(C)// Get the SF set by Proposition 2
  (2) K=|SF| //Number of features in SF
  (3) For i=1 to K //Find G by Definition 12
        $G_i = Get - PC(SF_i) \cup SF_i, SF_i \in SF$
      End
Step 2: Select features from representative sets
  (4) Divide G into K non-overlapping sets:
           $G_i \cap G_j = \emptyset, i \neq j$
  (5) Solve Eq. (3)
  (6) Return a best feature subset.
```

Figure 4. The SRS algorithm

In Step 1, we can get the representative sets by learning Bayesian networks. Instead of learning a complete Bayesian network among all features, we adopt a local Bayesian network learning strategy to discover the strongly relevant features of the class attribute (Proposition 2). Once we get the strongly relevant features, we use the same local learning technique to select the correlated features for each strongly relevant feature (Definition 12). Thus, as for the GET-PC function, we can use MMPC or HITON-PC which are both the state-of-the-art local Bayesian network learning algorithms (detailed descriptions in [1])[1]. Since both algorithms have very similar results, we employ HITON-PC as the GET-PC function in Step 1.

In Step 2, to solve Eq. (3), we employ a standard sparse group lasso using a least square loss function[2]. Instead of all possible sets (which could involve all features), in Step 2, the sparse group lasso method only needs to optimize over a small number of representative sets including the most informative features.

V. Experimental Results

A. Experimental Setup

We have chosen 16 benchmark datasets as described in Table 1. There are 5 datasets from the UCI machine learning repository (the first 5 datasets), 3 biomedical datasets (hiva, ovarian-cancer, and breast-cancer), 4 NIPS 2003 feature selection challenge datasets (arcene, dexter, dorothea, and madelon), and 4 public microarray datasets (the last 4 datasets) [21]. In our experiments, we treat those datasets in Table I as data in non-faithful distributions.

For the 4 NIPS 2003 challenge datasets and the spect dataset, we use the originally provided training and validation sets; for the 4 gene datasets we adopt the first 2/3 instances for training and the last 1/3 instances for testing; and for the rest datasets we use 10-fold cross-validation.

Table I Summary of the benchmark datasets.
(#F: Number of features, #I: Number of instances)

Dataset	#F	#I	Dataset	#F	#I
spect	22	267	madelon	500	2000
wdbc	30	569	colon	2000	62
spectf	44	267	prostate	6033	102
promoter	57	106	leukemia	7129	72
infant	86	5337	lung-cancer	12533	181
arcene	10000	100	breast-cancer	17816	286
dexter	20000	300	ovarian-cancer	2190	216
dorothea	100000	800	hiva	1617	4229

We use two classifiers, Knn and J48 provided by the Spider machine learning package[3]. Our comparative study uses five state-of-the-art Markov blanket filters, including IAMB [19], MMMB [20], PCMB [14], HITON-MB [1], and HITON-PC (only discovering parents and children of a target feature) [1], the state-of-the-art multiple Markov blanket discovery algorithm TIE* [17], and four well-established feature selection algorithms, FCBF [22], mRMR [15], SPSF-LAR [26], and MRF [3]. For parameter settings, λ_1 and λ_2 are both varied from [0.001, 0.1] with step 0.005 for SRS; and the significant level is set 0.01 for IAMB, MMMB, PCMB, HITON-PC and HITON-MB. All experiments were conducted on a computer with Inter(R) i7-2600 3.4GHz CPU and 12GB memory.

B. Comparison with HITON-PC, HITON-MB, and RES

Figures 5 and 6 summarize the classification errors of SRS against HITON-PC, HITON-MB and RES, using the Knn and J48 classifiers. RES (**RE**presentative **S**ets) means that we use the union of representative sets as a feature subset and calculate its classification errors. In both figures, points above the y = x diagonal are datasets for which SRS achieved lower classification errors than the competing algorithm. From Figures 5 to 11, we have three findings.

Firstly, from Figures 5 to 6, when we process data in non-faithful distributions, the Markov blanket selected by HITON-MB might not be an optimal solution to feature selection while SRS outperforms HITON-MB, HITON-PC and RES, especially on datasets with a small sample-to-feature ratio. Moreover, from Figure 11, on the number of selected features, SRS is also very competitive with HITON-MB and HITON-PC.

Secondly, in Figures 9 and 11, we can see that the selection of both parents and children of the class attribute may be enough instead of MBs since HITON-PC is superior to HITON-MB on both classification errors and the number of selected features on most datasets. Furthermore, HITON-PC is faster than HITON-MB and SRS since it only needs to discover parents and children. This validates Theorem 6 that in Markov blankets only parents and children are strongly relevant features.

Thirdly, Figure 10 illustrates that RES also gets excellent results on classification errors, even better than HITON-MB. Thus, we can conclude that representative sets contain sufficiently predictive features and we only need to focus on representative sets without an exhaustive search over all candidate MBs.

C. Comparison with Other Markov Blanket Filters

Figures 7 and 8 summarize the classification errors of SRS against IAMB, MMMB, and PCMB (points above the y = x diagonal are datasets for which SRS achieved lower classification errors than the competing algorithm). MMMB fails on the dorothea and leukemia datasets due to long running time (exceeding three days). Figures 7 and 8 show that SRS significantly outperforms those MB filters on the classification errors, especially using the Knn

[1] The codes of HITON_PC are available at http://www.dsl-lab.org/causal_explorer.

[2] The codes of sparse group lasso are available at http://www.public.asu.edu/~jye02/Software/SLEP/index.htm

[3] The Spider machine learning package in Matlab is available at http://people.kyb.tuebingen.mpg.de/spider/

classifier. As for the number of selected features, in Figure 12, SRS is also very competitive with its rivals since it considers not only the strongly relevant features but also their corresponding correlated features.

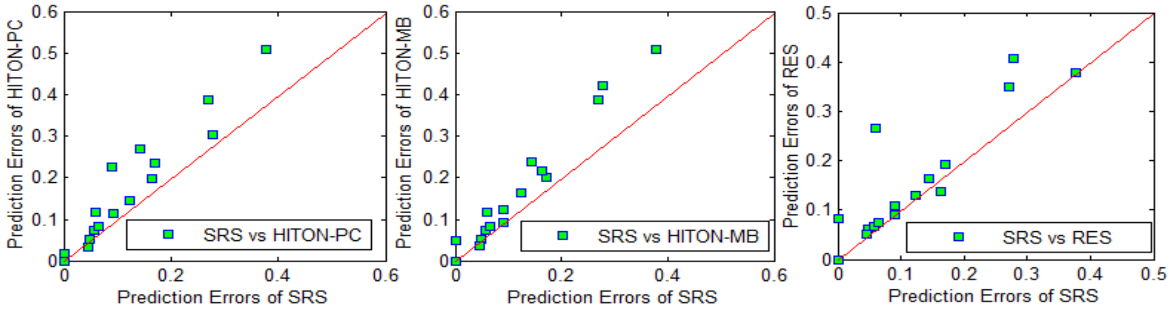

Figure 5. (Knn): Classification errors of SRS vs. HITON-PC, HITON-MB, and RES

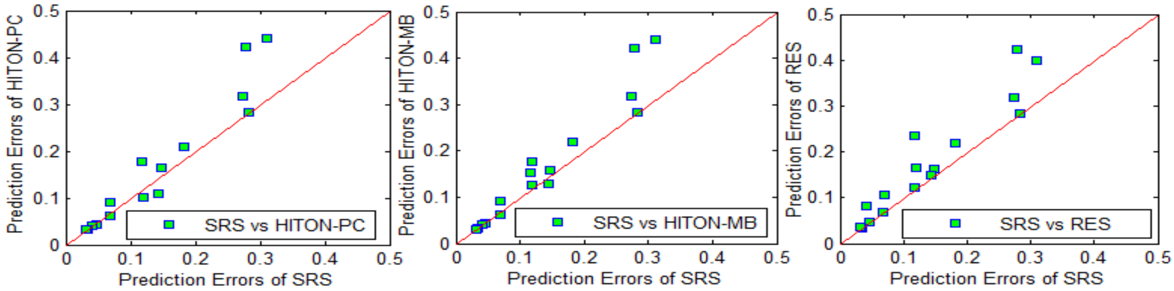

Figure 6. (J48): Classification errors of SRS vs. HITON-PC, HITON-MB, and RES

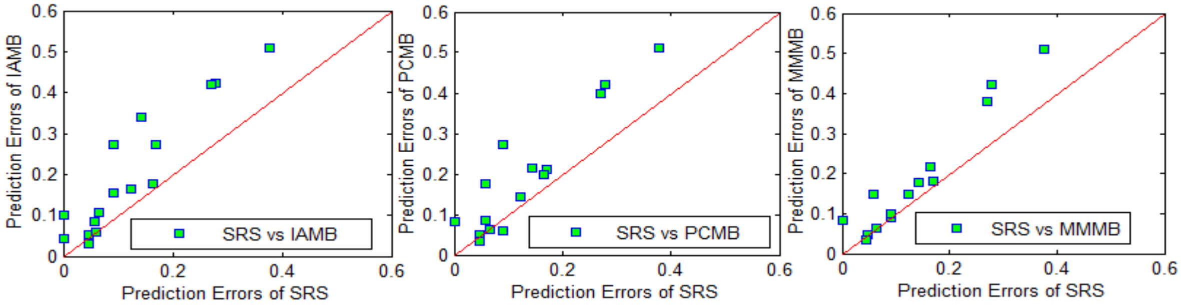

Figure 7. (Knn): Classification errors of SRS vs. IAMB, PCMB, and MMMB

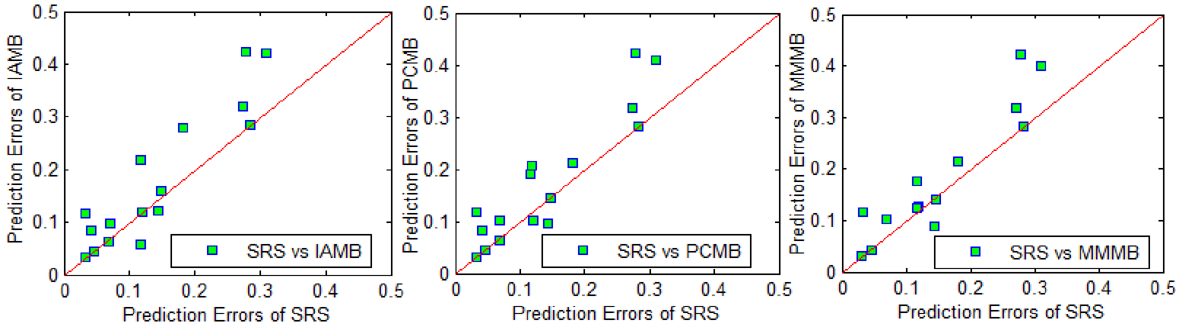

Figure 8. (J48): Classification errors of SRS vs. IAMB, PCMB, and MMMB

From Figure 13, on running time (in seconds), SRS is also very competitive with the other MB filters, even though it needs to consider not only the strongly relevant features but also their corresponding correlated features. We don't present the dorothea and leukemia datasets as MMMB fails on them while on colon, the running time is as follows: SRS: 63; HITON-MB: 2013; IAMB: 1, and PCMB: 1.

In summary, from Figures 5 to 13, our empirical study has indicated that when we process data in non-faithful distributions, MBs selected by the existing MB feature selection methods may not be an optimal feature subset, especially on datasets with a small sample-to-feature ratio while SRS can effectively and efficiently handle MB feature selection in real data. More importantly, with representative sets, SRS can efficiently find a best feature subset without an exhaustive search over an unknown space of the all MBs in each dataset.

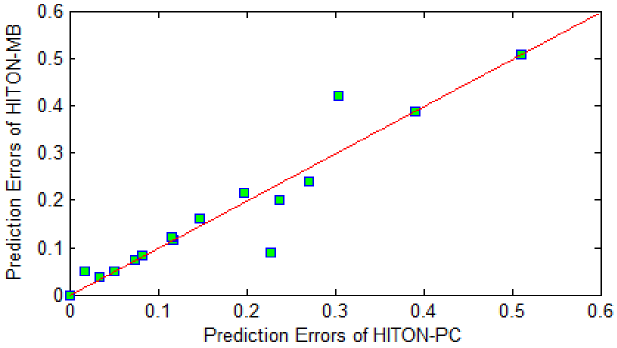

Figure 9. (Knn): Classification errors of HITON-PC vs. HITON-MB

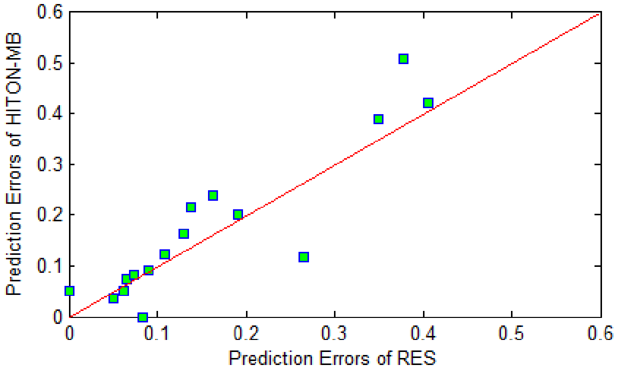

Figure 10. (Knn): Classification errors of HITON-MB vs. RES

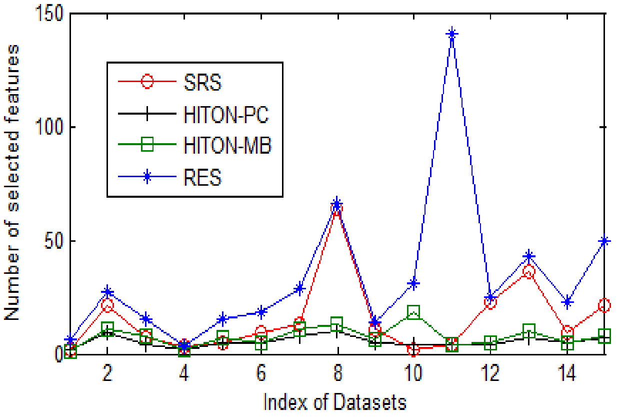

Figure 11. Number of selected features of SRS vs. HITON-PC, HITON-MB and RES

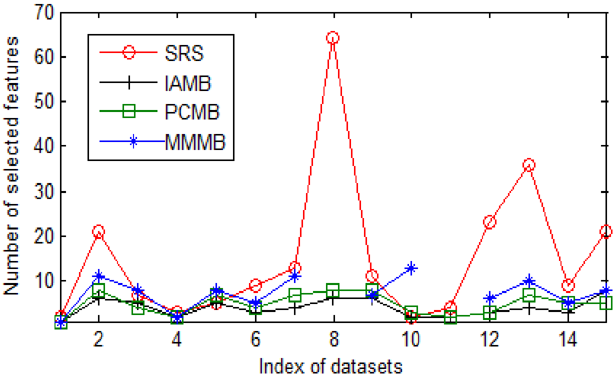

Figure 12. Number of selected features of SRS vs. IAMB, PCMB, and MMMB

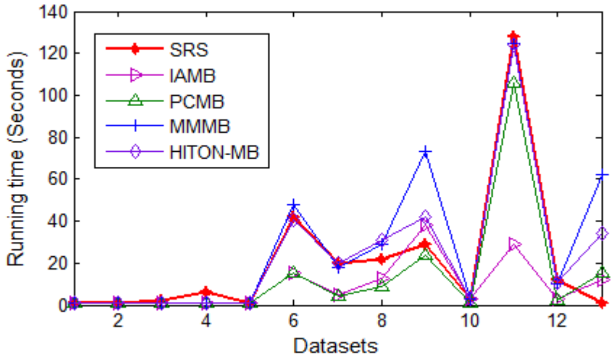

Figure 13. Running time (in seconds) of SRS against the other rivals (the labels of the x-axis from 1 to 13 denote the datasets in the left figure: 1. wdbc, 2. spectf, 3. infant, 4. promoter, 5. lung-cancer, 6. prostate, 7. arcene, 8. dexter, 9. madelon, 10. breast-cancer, 11. ovarian-cancer, 12. hiva, and 13. spect)

D. Comparison with the TIE algorithm*

In this section, we compare our SRS algorithm with the state-of-the-art multiple MB discovery algorithm, the TIE* algorithm which attempts to find all MBs in real data with non-faithful distributions. In our experiments, with the same parameter setting of the TIE* algorithm in [17], TIE* is parameterized with Semi-Interleaved HITON-PC as the base Markov blanket induction algorithm and a classification error as a criterion that verifies whether a new feature subset is a Markov blanket of the class attribute. The parameter alpha of Semi-Interleaved HITON-PC is set 0.05. We selected the Markov blanket with the lowest classification error from all of the MBs discovered by TIE*. In the following figures, we don't plot the errors of the ovarian-cancer dataset for SRS and TIE*, since TIE* failed on this dataset due to long running time (exceeding three days).

From Figures 14 to 15, we can see that SRS outperforms TIE* on most of the datasets. Why is SRS superior to TIE*? The possible explanation is that TIE* simply selects one feature from a set of strongly correlated features while SRS might pick out more features from a group of strongly correlated features, and this might be beneficial to reduce classification error. This also explains why SRS selects more features than TIE* as shown in Table II. For example, on the spectf dataset, SRS gets four group features {5,6,32}, {14,16,24,25,26}, {30,39,40}, and{15,41,42,43}, and features in each group are strongly correlated. SRS selects seven features {5, 32, 24, 26, 30, 40, 15} from those groups while the MB selected by TIE* only contains two features 30 and 42 which attain the lowest errors among all Markov blankets. On Knn and J48, SRS gets the classification errors 16.42% and 11.94% respectively, while TIE* attains the errors 20.5% and 16.04% respectively.

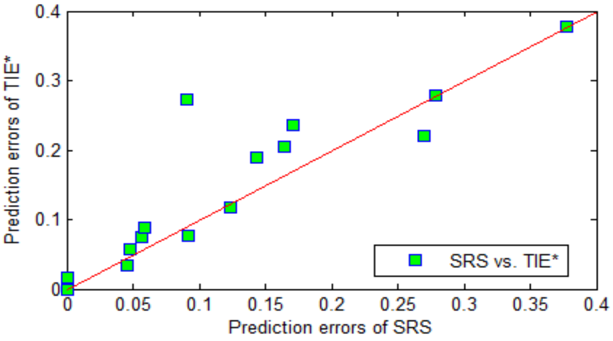

Figure 14. (Knn): Classification errors of SRS vs. TIE*

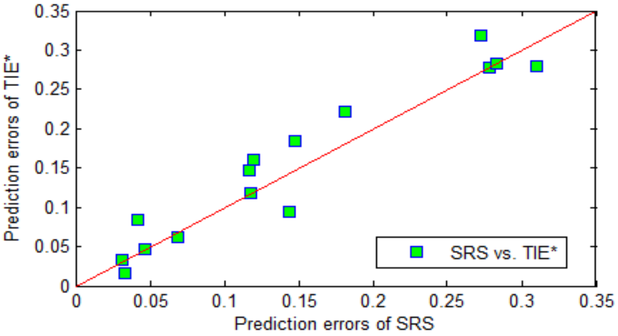

Figure 15. (J48): Classification errors of SRS vs. TIE*

TABLE II NUMBERS OF SELECTED FEATURES AND RUNNING TIME

Dataset	# Selected features		Running time	
	SRS	TIE*	SRS	TIE*
spect	2	1	1	1
wdbc	21	8	1	57
spectf	7	2	1	2
promoter	3	3	6	1
infant	5	2	2	128
arcene	9	3	20	1292
dexter	13	4	22	190
dorothea	64	5	848	10173
madelon	11	5	29	121
colon	101	2	63	4
prostate	2	2	42	26
leukemia	4	2	468	31
lung-cancer	23	3	1	361
breast-cancer	36	6	3	198080
hiva	21	4	12	254
ovarian-cancer	9	/	128	/

From Table II, we can see that SRS is much faster than TIE*, especially on the arcene, dorothea and breast-cancer datasets. On the ovarian-cancer datasets, TIE* failed due to long running time (exceeding three days). But why on the colon and leukemia datasets, is TIE* faster than SRS? The main reason is that Semi-HITON-PC employed in TIE* is faster than HITON-PC used in SRS.

In summary, instead of an exhaustive search for all Markov blankets, the discovery of a best Markov blanket from representative sets is not only more effective but also more efficient than TIE*.

E. Comparison with Other Feature Selection Methods

Figures 16 and 17 present the classification errors of SRS against two well-established feature selection algorithms, FCBF and mRMR, and two state-of-the-art algorithms, SPSF-LAR, and MRF.

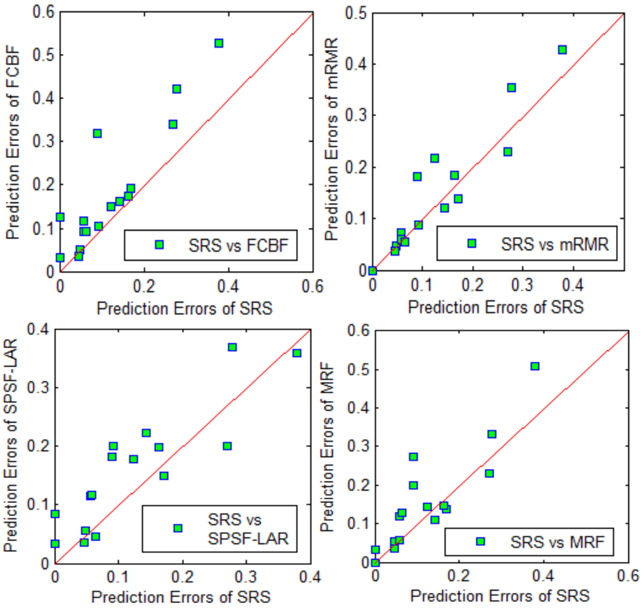

Figure 16. (Knn): Classification errors of SRS vs. FCBF, mRMR, SPSF-LAR and MRF

Since SRS selects no more than 65 features to get the lowest classification error on all 16 datasets, we set the parameter k for the SPSF-LAR, MRF, and mRMR methods from 1 to a maximum number of 60, respectively. For 5 UCI datasets, we use the feature subset whose size ranges from 1 to 15 and choose the lowest classification error rate achieved by Knn and J48 while for the remaining 11 high-dimensional datasets, we use the top 5, 10, 15, ..., 60 features selected by each algorithm.

From Figures 16 to 17, we can see that SRS often outperforms the other rivals by using Knn while it produces significantly better results than the other algorithms by using J48.

VI. CONCLUSION

In this paper, we explored Markov blanket feature selection by dealing with data in non-faithful distributions.

To tackle this issue, we extended the concept of Markov blankets and proposed the concept of representative sets. With representative sets, we presented the SRS algorithm for Markov blanket feature selection by employing a standard sparse group lasso. The experimental results have shown that the SRS selector outperforms both state-of-the-art Markov blanket feature selectors and other well-established feature selection methods on real datasets.

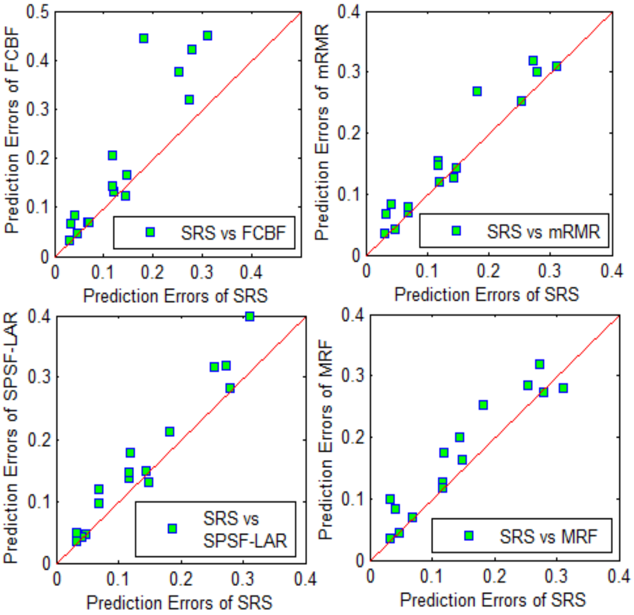

Figure 17. (J48): Classification errors of SRS vs. FCBF, mRMR, SPSF-LAR and MRF

Acknowledgment

This work was 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 (61305064, 61070131, 61175033, and 61229301).

References

[1] 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.

[2] G. Brown, A. Pocock, M. Zhao, and M. Luján. (2012) Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. Journal of Machine Learning Research, 12, 27-66.

[3] Q. Cheng, H. Zhou, and J. Cheng. (2011) The Fisher-Markov Selector: Fast Selecting Maximally Separable Feature Subset for Multiclass Classification with Applications to High-Dimensional Data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(6), 1217-1233.

[4] J. Friedman, T. Hastie, and R. Tibshirani. (2010) A Note on the Group Lasso and a Sparse Group Lasso. Arxiv preprint arXiv:1001.0736.

[5] I. Guyon and A. Elisseeff. (2003) An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3, 1157-1182

[6] 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.

[7] D. Koller and M. Sahami. (1996) Toward Optimal Feature Selection. ICML'96, 284-292.

[8] R. Kohavi and G. H. John. (1997) Wrappers for Feature Subset Selection. Artificial Intelligence, 97, 273-324.

[9] M. A. Hall.(2000) Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. ICML'00, 359–366.

[10] H. Liu and L. Yu. (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions onKnowledge and Data Engineering,17(4), 491-502.

[11] D. Margaritis and S. Thrun. (2000) Bayesian Network Induction via Local Neighborhoods. In Advances in Neural Information ProcessingSystems 1999. Denver, Colorado, USA: The MIT Press.

[12] R. E. Neapolitan (2004) Learning Bayesian Networks. Prentice Hall series in artificial intelligence. Pearson Prentice Hall, Upper Saddle River, NJ.

[13] J. Pearl (1988) Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, CA.

[14] 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.

[15] H. Peng, F. Long, and C. Ding. (2005) Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1226-1238.

[16] J. Shen, L. Li, and W-K. Wong. (2008) Markov Blanket Feature Selection for Support Vector Machines. AAAI'08, 696-701.

[17] A. Statnikov, N. Lytkin, J. Lemeire and C. F. Aliferis. (2013) Algorithms for Discovery of Multiple Markov Boundaries. Journal of Machine Learning Research 14, 499-566.

[18] R. Tibshirani. (1996) Regression Shrinkage and Selection via the Lasso. J. Roy. Stat. Soc. B, 58(1):267-288.

[19] I. Tsamardinos and C. F. Aliferis. (2003) Towards Principled Feature Selection: Relevancy, Filters and Wrappers. AI & Statistics'03.

[20] 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.

[21] L. Yu, C. Ding, and S. Loscalzo. (2008) Stable Feature Selection via Dense Feature Groups. KDD'08, 803-811.

[22] L. Yu and H. Liu. (2004) Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research, 5: 1205-1224.

[23] K. Yu, X. Wu, W. Ding, H. Wang, and H. Yao. (2011) Causal Associative Classification. ICDM'11,914 – 923.

[24] K. Yu., W. Ding, H. Wang, and X. Wu. (2013) Bridging Causal Relevance and Pattern Discriminability:Mining Emerging Patterns from High-Dimensional Data. IEEE Transactions on Knowledge and Data Engineering. In press.

[25] M. Yuan and Y. Lin. (2006) Model Selection and Estimation in Regression with Grouped Variables. J. Roy. Stat.Soc. B, 49-67.

[26] Z. Zhao, L. Wang, H. Liu, and J. Ye. (2013) On Similarity Preserving Feature Selection. IEEE Transactions on Knowledge and Data Engineering, 25(3), 619-632.