Markov Blanket Feature Selection using Representative Sets

Kui Yu, Xindong Wu*, *Fellow*, *IEEE*, Wei Ding, *Senior Member*, *IEEE*, Yang Mu, and Hao Wang

***Abstract*— It has received much attention in recent years to use Markov blankets in a Bayesian network for feature selection. The Markov blanket of a class attribute in a Bayesian network is a unique yet minimal feature subset for optimal feature selection if the probability distribution of a data set can be faithfully represented by this Bayesian network. However, if a data set violates the faithful condition, Markov blankets of a class attribute may not be unique. To tackle this issue, in this paper, we propose a new concept of representative sets, and then design the SGAI (Selection via Group Alpha-Investing) algorithm to perform Markov blanket feature selection with representative sets for classification. Using a comprehensive set of real data, our empirical studies have demonstrated that SGAI outperforms the state-of-the-art Markov blanket feature selectors and other well-established feature selection methods.**

***Index Terms*—Feature Selection, Representative Sets, Markov Blankets, Bayesian Networks**

I. Introduction

FEATURE selection is to select a subset of relevant features from an original feature space to improve the performance of prediction models in terms of their accuracy, efficiency, and model interpretability [1, 2, 14, 26]. With respect to a class attribute, an input feature can be classified into a strongly relevant, irrelevant, redundant, or non-redundant feature [11, 29]. The task of feature selection is to choose a feature subset including strongly relevant and non-redundant features.

Koller and Sahami [10] was the first to introduce Markov blankets into feature selection for removing irrelevant or redundant features. A Markov blanket (boundary) was first invented in a Bayesian network by Pearl [17], and is defined as that in a faithful Bayesian network, for every node (feature) X, its Markov blanket is the set of parents, children and spouses (parents of the children of X), as shown in Figure 1.

- *Kui Yu is with the School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, 5095, Australia. E-mail: Kui.Yu@unisa.edu.au*
- *Xindong Wu (*Corresponding author) is with the School of Computing and Informatics, University of Louisiana, Lafayette, Louisiana 70504, USA. E-mail: xwu@louisiana.edu.*
- *Wei Ding and Yang Mu are with the Department of Computer Science, University of Massachusetts Boston, Boston, 02125, USA. E-mail: {yangmu, ding}@cs.umb.edu.*
- *Hao Wang is with the Department of Computer Science, Hefei University of Technology, Hefei, 230009, China. E-mail: jsjxwangh@hfut.edu.cn.*

Technically, a Bayesian network is presented by a directed acyclic graph $\mathbb{G}$ and a joint probability distribution $\mathbb{P}$ over a variable set F. $\mathbb{G}$ is faithful to $\mathbb{P}$ if and only if every independence present in $\mathbb{P}$ is entailed by $\mathbb{G}$ and the Markov condition. A joint probability distribution $\mathbb{P}$ is faithful if and only if there exists a directed acyclic graph $\mathbb{G}$ such that $\mathbb{G}$ is faithful to $\mathbb{P}$ [17]. If $\mathbb{G}$ and $\mathbb{P}$ are faithful to each other, the Bayesian network represented by $\mathbb{G}$ and $\mathbb{P}$ is said to satisfy the faithful condition. In principle, a Markov blanket of feature X is a minimal set of features with the following property [10]: for every feature Y not in the Markov blanket, Y and X are conditionally independent given the Markov blanket of X.

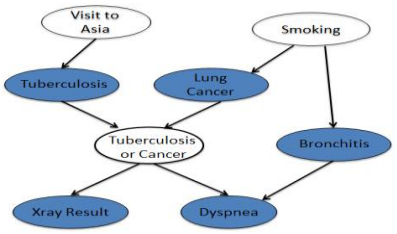

Fig.1. The Markov blanket (in blue) of the node of "Tuberculosis or Cancer" in the Bayesian network of Asia [17].

Tsamardinos and Aliferis [23] bridged the gap between the concepts of feature relevance in feature selection and Markov blankets in a Bayesian network for classification, and transferred the feature selection problem to the problem of discovery of the Markov blanket of a class attribute in a faithful Bayesian network. They theoretically proved that if a probability distribution can be faithfully represented by a Bayesian network (the faithful condition), then the Markov blanket of a class attribute in the Bayesian network is not only unique, but also the set of strongly relevant features as defined by Kohavi and John [11]. Tsamardinos and Aliferis [23] proposed the IAMB (Incremental Association-Based Markov Blanket) algorithm for optimal feature selection. The IAMB algorithm can return a Markov blanket of any target node in a Bayesian network without learning a complete Bayesian network, even with hundreds of thousands of features. Thus, the discovery of Markov blankets from Bayesian networks for feature selection has attracted much attention [3, 18, 31]. More recent variations of IAMB include PCMB (Parent-Children Markov Blanket) [18], MMMB (Max-Min Markov Blanket) [24], and HITON-MB [3].

All the existing studies on Markov blankets typically assume that a data distribution and an underlying Bayesian network which models the domain are faithful to each other, in order to guarantee a target node in a Bayesian network has a unique Markov blanket. However, many data sets from real-world applications may violate the faithful condition, and this makes Markov blankets of a target feature not unique. For instance, in Figure 2, if the Bayesian network is parameterized such that X and Y carry equivalent information about T, then there may be two Markov blankets of T: { X;A} and { Y;A} [21]. We now further explain the existence of multiple Markov blankets using two real examples.

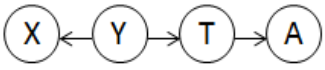

Fig. 2. A symbolic example of multiple Markov blankets

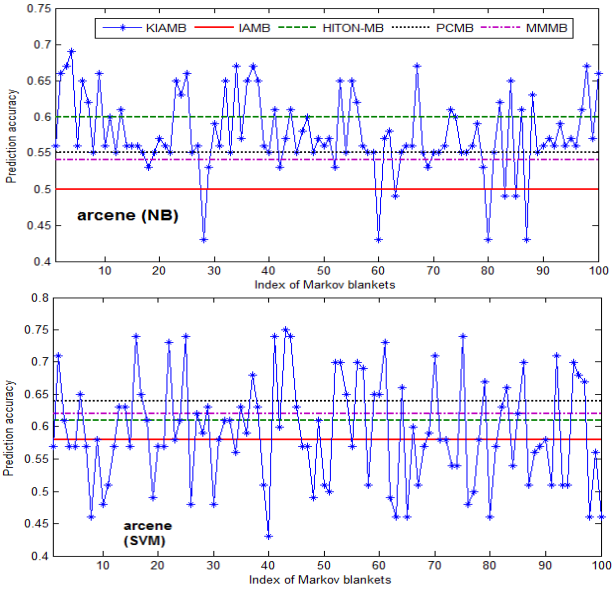

Fig.3. Accuracy on the arcene data set by exploring the non-faithful problem

In analysis of real-world high-throughput molecular data, due to different gene or biomarker sets performing almost equally well in terms of prediction accuracy of phenotypes, it is a ubiquitous phenomenon of the existence of multiple Markov blankets in these data sets [21].

Another real-world application example is of catastrophic flood forecasting [25]. Due to spatial and temporal adjacency, many meteorological predictor variables contain equivalent information for flood prediction, because these predictor variables are not physically independent of one another. For example, when there is a dropping in sea level pressure (anomalously low sea level pressure is an important characteristic of atmospheric regimes, which may lead to extreme precipitation clusters), at the same location there must also be convergence of the winds near the surface and divergence of the winds at the top of the atmosphere. In this case, the low level winds, high level winds, and vertical motions in the same location are considered equivalent meteorological predictor variables.

Figure 3 investigates this non-faithful problem using a real data set of the arcene (a cancer benchmark data set with 100 instances and 10,000 features) used by the NIPS 2003 feature selection challenge. In Figure 3, the state of the art Markov blanket feature selection methods IAMB, HITON-MB, PCMB and MMMB can only discover a single Markov blanket. Different from those four algorithms, KIAMB [18] can find multiple sets of Markov blankets, by employing a stochastic search heuristic that repeatedly disrupts the order in which features are selected for inclusion into a Markov blanket with the probability p at each round, 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 the Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers, Figure 3 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.

From Figure 3, we can see that the Markov blankets identified by IAMB, HITON-MB, PCMB, and MMMB may not be the feature subsets that maximize the prediction accuracies, compared to the ones discovered by KIAMB.

The main challenges are as follows to deal with Markov blanket feature selection under the non-faithful condition.

Challenge 1. *The number of Markov blankets varies from different data sets and may grow exponentially with respect to the number of features in the underlying Bayesian network.*

Challenge 2. *It is difficult to efficiently identify a best feature subset from multiple Markov blankets.* It may be time-consuming or computationally prohibitive to find a best Markov blanket that maximizes the predictive power for classification from a large, or even exponential number of Markov blankets.

In this paper,

(1) We propose a new concept of representative sets, instead of Markov blankets. We focus on Markov blanket feature selection with representative sets, instead of an exhaustive search over a large number of Markov blankets in real data.

(2) We propose the SGAI (Selection via Group Alpha-Investing) algorithm for selecting feature subsets from representative sets. Compared to the SRS algorithm proposed in our preliminary version [31], the SGAI algorithm avoids computing the regularization parameters of SRS.

(3) We have conducted comprehensive experiments using 16 data sets and 10 state-of-the-art feature selectors to validate the effectiveness and efficiency of our method.

II. Related Work

Feature selection aims to reduce the computational complexity without performance degradation by removing irrelevant and redundant features [11, 27]. 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 [19] and the FCBF (Fast Correlation Based Filter) algorithm

[29]. Recently, Cheng et al. [6] 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). Zhao et al. [31] proposed a framework to unify different criteria for handling feature redundancies. Brown et al. [4] unified almost two decades of research on information theoretic feature selection to an optimization of conditional likelihoods.

Tsamardinos and Aliferis [23] associated Markov blankets in faithful Bayesian networks with strongly relevant features defined by Kohavi and John [11], and then transferred the feature selection task to the discovery of Markov blankets in a faithful Bayesian network. Since then, Markov blanket feature selection as an emerging successful class of filter methods has attracted much attention [3, 18, 30].

Margaritis and Thrun [14] invented the first yet sound algorithm, the GS (Grow/Shrink) algorithm, with the intent to find Markov blankets for the purpose of speeding up global Bayesian network learning. The GS algorithm requires exponential number of instances to the size of the Markov blanket, thus impractical for many real data sets.

To conquer this drawback of the GS algorithm, Tsamardinos and Aliferis [23] 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 the method is more sample efficient than GS. However, the IAMB algorithm still requires a sample size exponential in the size of a Markov blanket. Thus, HITON-MB and MMMB were introduced to mitigate the problem of data inefficiency. Different from GS and IAMB, HITON-MB [3] and MMMB [24] 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 [1] and MMPC were introduced [24]. Following the idea of MMMB, PCMB [18] was also proposed to conquer the data inefficiency problem.

All these algorithms are well-established only for selecting a single Markov blanket under the assumption that probability distribution can be faithfully represented by an underlying Bayesian network. A naïve approach for handling Markov blanket feature selection under non-faithful conditions 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 dimensionality, and features in each cluster do not indicate they are correlated in terms of feature relevance [13, 28]. Peña et al. [18] proposed a stochastic Markov blanket algorithm based on IAMB, called KIAMB, which involves running multiple times initialized with a random seed. But we do not know how many times the KIAMB algorithm need to run to get an optimal feature subset for feature selection. Recently, among the most notable advances in the field is that Statnikov et al. [21] proposed the TIE* (Target Information Equivalence) algorithm that can discover all Markov blankets under non-faithful conditions and outperforms KIAMB. But TIE* preferentially discovers multiple Markov blankets for causal discovery for improving the causal induction mechanisms without missing causative variables and is not yet customized for feature selection. Furthermore, TIE* may be computationally expensive when the number of Markov blankets grows exponentially with respect to the number of features in the network.

III. Notations and Definitions

In the following sections, we will introduce Markov blankets, Bayesian networks, and Markov blanket feature selection. Table 1 summarizes and explains the notations and symbols used in this paper.

A. Markov Blankets in Feature Selection

To characterize feature relevance, Kohavi and John [11] classified input features into three disjoint feature sets, strongly relevant, weakly relevant, and irrelevant subsets. Later Yu and Liu [29] divided weakly relevant features into redundant features and non-redundant features.

TABLE 1
SUMMARY ON MATHEMATICAL NOTATIONS

Notations	Mathematical Meanings
F	an input feature set
F_i, X, Y, T	a single feature (attribute)
f_i	a disrete value of F_i taking
$F - \{F_i\}$	a feature subset excluding F_i
S, V, Z, W	a feature subset within F
C	a class attribute
M_i	a Markov blanket of feature F_i
$\mathbb{G}$	a directed acyclic graph over F
$\mathbb{P}$	a discrete joint probability distribution over F
$\|.\|$	$\|F\|$ returns the number features in F
$P(.\|.)$	$P(F_i\|F_j)$ denotes the posterior probability of F_i conditioned on F_j
$\Im$	a seed set of a class attribute
$Ind\ (X,Y\|Z)$	X and Y are conditionally independent given Z
$\wp(F_i)$	a set of correlated features of F_i
$Pa(.)$	$Pa(F_i)$ denotes the set of parents of F_i
$Ch(.)$	$Ch(F_i)$ denotes the set of children of F_i
$\mathcal{L}(\cdot)$	a loss function
$\mathcal{R}$	a set of representive sets
$\mathcal{R}_i$	a representative set
$\Omega(.)$	a regularization term
β	a set of coefficient vectors
β_i	a coefficient vector corresponding to $\mathcal{R}_i$
$W(\mathcal{R}_i)$	the weight of $\mathcal{R}_i$

Definition 1 (Conditional Independence) Feature $F_i \in F$ is conditionally independent of $F_k \in F$ $(\mathrm{i} \neq \mathrm{k}, \mathrm{i}, \mathrm{k} = 1 \ldots \mathrm{n})$ conditioned on $S \subseteq F - \{F_i \cup F_k\}$ if

$$P(F_i|F_k, S) = P(F_i|S). \quad (1)$$

Definition 2 (Markov Blanket) [10] A Markov blanket of feature F_i, denoted as $M_i \subset F - \{F_i\}$ is a minimal set of features that makes every other feature independent of F_i given M_i, that is,

$$\forall F_k \in F - (M_i \cup \{F_i\}) \ \ s.t.\ P(F_i|M_i, F_k) = P(F_i|M_i). \quad (2)$$

Definition 3 (Redundant Feature) A feature F_i is redundant and hence should be discarded from F, if it has a Markov blanket within F.

According to Definition 2, a Markov blanket of a feature F_i subsumes the information what F_i has about a class attribute, while the Markov blanket of the class attribute carries information what all of the other features have about the class attribute. In other words, the Markov blanket of a class attribute is the optimal feature subset which should contain strongly relevant and non-redundant features [10, 29].

B. Markov Blankets in Bayesian Networks

Definition 4 (Bayesian network) [17] 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.

With the Markov condition, a Bayesian network encodes the joint probability $\mathbb{P}$ over F and decomposes $\mathbb{P}$ into a product of the conditional probability distributions over each node F_i given its parents $Pa(F_i)$ in $\mathbb{G}$. The joint probability $\mathbb{P}$ is written as follows.

$$\mathbb{P}(F_1, F_2, \dots, F_n) = \prod_{i=1}^{n} \mathbb{P}(F_i | Pa(F_i)) \quad (3)$$

Theorem 1. [17] *Let S, V, Z, and W be any four disjoint 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:*

$$Ind(S, V|Z \cup W) \text{ and } Ind(S, W|Z \cup V) \Rightarrow Ind(S, (V \cup W)|Z) \quad (4)$$

where the notation *Ind* (*S,V*|*Z*) denotes that two sets of features *S* and *V* are conditionally independent given a set of features *Z* over the joint probability distribution $\mathbb{P}$.

Theorem 2. [17] *If a joint probability distribution* $\mathbb{P}$ *over a feature set F satisfies the intersection property, then for each* $F_i \in F$*, there exists a unique Markov blanket of* F_i.

Definition 5 (Faithfulness) [3, 17] Given a Bayesian network $< F, \mathbb{G}, \mathbb{P} >$, $\mathbb{G}$ is faithful to $\mathbb{P}$ over F iff every independence present in $\mathbb{P}$ is entailed by $\mathbb{G}$ and the Markov condition. $\mathbb{P}$ is faithful iff there exists a directed acyclic graph $\mathbb{G}$ such that $\mathbb{G}$ is faithful to $\mathbb{P}$.

Theorem 3. [17] *If* $\mathbb{P}$ *is faithful to* $\mathbb{G}$*, then* $\mathbb{P}$ *satisfies the intersection property.*

With Theorems 2 and 3, we give the definition of Markov blankets in a faithful Bayesian network.

Definition 6 (Markov Blanket) [23] In a faithful Bayesian network, for every node F_i, its Markov blanket is unique with the set of parents, children and spouses of F_i.

C. Markov Blanket Feature Selection

Koller and Sahami [10] is the first to introduce Markov blankets to feature selection for removing irrelevant and redundant features. They proposed a prototype algorithm for Markov Blanket feature selection: let F be an input feature set, if $\exists F_i \in F$, there exists a Markov blanket of F_i, F_i can be removed from F [10]. However, it was not guaranteed to discover the actual Markov blanket and nor could it be scaled to high dimensionality [3]. Tsamardinos and Aliferis [23] proposed Theorem 4 below to link feature relevance and Markov blanket in Bayesian networks for feature selection.

Theorem 4. *A feature is strongly relevant to C, iff it belongs to the Markov blanket of C in a faithful Bayesian network.*

Theorem 4 confirms that we can transfer the feature selection problem into the task of the discovery of the Markov blanket of a class attribute in a faithful Bayesian network.

IV. Selection via Representative Sets

A. Representative Sets

When a data set does not satisfy the faithful condition, it may contain multiple Markov blankets of a target feature [18, 21]. Figure 3 illustrates that some feature sets (redundant features) discarded by the existing Markov blanket algorithms actually carry a stronger predictive ability than the selected Markov blankets. Feature redundancy is usually defined by means of feature correlation [5, 29], and thus we call those redundant yet discarded features as correlated features with regard to the selected features, which are defined as follows.

Definition 7 (Correlated Feature) F_k is called a correlated feature of F_i, if $\forall S \subseteq F, \{F_k, F_i\} \notin S, P(F_i|\mathrm{S}, F_k) \neq P(F_i|\mathrm{S})$.

With correlated features, the feature space of possible Markov blankets may consist of two parts: features in a Markov blanket and their corresponding correlated features. Clearly, if we get those two parts of features, it is more practical to select a best Markov blanket that maximizes the predictive power for classification from this reduced feature space than from an entire feature space. With those observations, we extend the concept of Markov blankets, and then propose the representative sets as the following to combat Markov blanket feature selection under the non-faithful condition.

Definition 8 (Seed Set) A seed set $\mathfrak{I}$ is defined as the Markov blanket of a class attribute in a faithful Bayesian network.

Definition 9 (Representative Set) We define $\mathcal{R} = \{\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_K\}$ as K representative sets with respect to K features within $\mathfrak{I}$, where each representative set $\mathcal{R}_i = \{F_i \cup \wp(F_i)\}$ with respect to F_i, F_i belongs to $\mathfrak{I}$, and $\wp(F_i)$ is a set of correlated features with respect to F_i.

Different from a Markov blanket, each member in a representative set $\mathcal{R}_i$ is not a single feature any more, but a feature set consisting of a feature $F_i \in \mathfrak{I}$ and $\wp(F_i)$. As for $\mathfrak{I}$,

we can employ an existing single Markov blanket discovery algorithm to get it. The key problem is how to determine $\wp(F_i)$, $F_i \in \Im$? Clearly, using Definition 7 to directly search for correlated features, it is expensive, or even prohibitive since the number of S is exponential in the dimensionality of a data set.

To calculate $\wp(F_i)$, using Bayesian networks, we give our main ideas using the following propositions and definitions. Given a Bayesian network, assuming $Pa(F_i)$ and $Ch(F_i)$ are the set of parents and children of F_i, respectively. By the Markov condition in Definition 4, we get the following.

Proposition 1. *In a Bayesian network* $< F, \mathbb{G}, \mathbb{P} >$*, if there exists a direct edge between* $F_i \in F$ *and* $F_k \in F$*, then* F_i *and* F_k *are directly correlated to each other.*

Proof. By the Markov condition in Definition 4, if node F_i belongs to $Pa(F_k)$, then F_i is independent of any subset of its non-descendant nodes conditioned on $Pa(F_i)$, and vice versa. Accordingly, F_i and F_k are conditionally dependent conditioned on any subsets within the remaining nodes in a Bayesian network. By Definition 7, the proposition is proved.

Proposition 1 illustrates that two nodes directed linked by an edge in a Bayesian network must be directly correlated to each other. Thus, the features directly correlated to F_i, that is, $\wp(F_i)$, can be defined in Proposition 2.

Proposition 2. In a Bayesian network $< F, \mathbb{G}, \mathbb{P} >$, $\wp(F_i) = \{Pa(F_i) \cup Ch(F_i) | F_i \in F\}$.

For a representative set $\mathcal{R}_i$ of C with respect to F_i, $F_k \in \wp(F_i)$ satisfies the following.

$$\exists S \in \Im, F_i \in S, \text{ s.t. } P(C|S, F_k) = P(C|S) \quad (5)$$

Eq.(5) means that the features in $\wp(F_i)$ are masked by some feature subsets in $\Im$ while running the existing single Markov blanket algorithms. Clearly, given $\Im = \{Pa(C) \cup Ch(C)\}$, there must exist a subset S within $\Im$ in order to make Eq.(5) hold. We get Proposition 3 as follows.

Proposition 3. Given a Bayesian network, the representative set $\mathcal{R}_i$ of C with respect to F_i is that $\mathcal{R}_i = \{F_i \cup \wp(F_i) | F_i \in Pa(C) \cup Ch(C)\}$.

Proposition 3 denotes that a representative set $\mathcal{R}_i$ includes a parent or a child of the class attribute C, and the set of parents and children of this parent or child. We do not treat a spouse in a Markov blanket and its parents and children as a representative set because a spouse has already been selected in a representative set. For instance, Figure 4 shows the four representative sets related to node "T" in the four red groups, considering node "T" as the class attribute. We can see that spouse "M" is included in one of those representative sets.

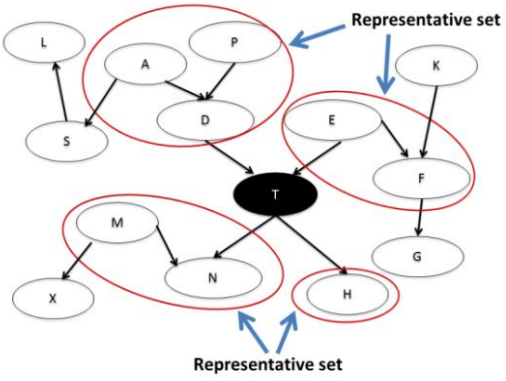

Fig.4. Representative sets related to node "T"

B. Selecting Features from Representative Sets

1) Problem Formulation

With representative sets, our problem becomes how to extract a best subset from representative sets? Different from a single feature set, to handle multiple representative sets, we need to learn 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 C is measured by a loss function $\mathcal{L}(\cdot)$, and $\mathcal{R} = \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_K\}$ represents K representative sets with respect to K features within $\Im$. Each element $\mathcal{R}_i$ (i = 1 ... K) in $\mathcal{R}$ incorporates the i[th] feature in $\Im$ and its corresponding correlated features.

The first step of our solution is to identify which representative sets have the most predictive power to *C*, and we formulate this step as follows.

$$\beta^* = arg\ min_\beta\ \mathcal{L}(\beta, \mathcal{R}, C) + \lambda_1 \sum_{i=1}^{K} \Omega_1(\beta_i), \quad (6)$$

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 $\mathcal{R}_i$, 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 sets, and if β_i=0, then the i[th] set will be excluded entirely.

The second step is to calculate the feature(s) to be chosen from each selected representative set, which is expected to contribute the most predictive power to *C*. The objective function in Eq. (6) is then further formulated as,

$$\beta^* = arg\ min_\beta\ \mathcal{L}(\beta, \mathcal{R}, C) + \lambda_1 \sum_{i=1}^{K} \Omega_1(\beta_i) + \lambda_2\ \Omega_2(\beta), \quad (7)$$

where $\Omega_2(\beta)$ penalizes the complexity of β. The parameter λ_2 adjusts the individual feature coefficient in 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.(7), in our preliminary version, we adopted the sparse group lasso approach and proposed the SRS (**S**election via **R**epresentative **S**ets) algorithm [31]. It is a nontrivial task to select decent λ_1 and λ_2 values on each data set for SRS to maximize the predictive power. In this paper, we propose a

new solution to solve Eq.(7) to avoid computing $\succ_1$ and $\succ_2$.

2) Solving Eq.(7) with Group Alpha-investing

How to assess the predictive power of each representative set. We assign a weight to each representative set $\mathcal{R}_i$ which controls the predictive power of $\mathcal{R}_i$ to *C*, and the weight of each representative set $\mathcal{R}_i$ is calculated in the two cases below.

Case 1: Calculating the weight of $\mathbf{R_i}$ in discrete cases

For discrete training data, we adopt the symmetrical uncertainty (*SU*) [29] to compute the weight of $\mathcal{R}_i$, and *SU* is defined as

$$SU(F_i, F_j) = 2\left[\frac{IG(F_i|F_j)}{H(F_i)+H(F_j)}\right], F_i \in \mathcal{R}_i, F_j \in \mathcal{R}_i, \quad (8)$$

where $IG(F_i|F_j)$ is information gain [20], and is computed by

$$IG(F_i|F_j) = H(F_i) - H(F_i|F_j), \quad (9)$$

where $H(F_i)$ is the entropy of F_i that is defined as

$$H(F_i) = -\sum_{f_i \in F_i} P(f_i) log_2(P(f_i)) \quad (10)$$

and $H(F_i|F_j)$ is the entropy of F_i after observing values of another feature F_j, which is defined as

$$H(F_i|F_j) = -\sum_{f_j \in F_j} P(f_j) \sum_{f_i \in F_i} P(f_i|f_j) log_2\left(P(f_i|f_j)\right). \quad (11)$$

We employ the *SU* to represent the relationships between features since it can compensate for information gain bias toward features with lots of values. The values of *SU* is restricted to the range [0, 1]. With the *SU*, we define the weight of $\mathcal{R}_i$:

$$W(\mathcal{R}_i) = \sum_{j=1}^{|\mathcal{R}_i|} SU(C, F_j). \quad (12)$$

Case 2: Calculating the weight of $\mathbf{R_i}$ in continuous cases

In continuous cases, we employ the partial correlations between features to compute the weight of a representative set. The correlation coefficient $\rho_{(X_i,Y_i)}$ between two variables X_i and Y_i is calculated as:

$$\rho_{(X_i,Y_i)} = \frac{\sum_i (X_i - \overline{X_i})(Y_i - \overline{Y_i})}{\sqrt{\sum_i (X_i - \overline{X_i})^2}\sqrt{\sum_i (Y_i - \overline{Y_i})^2}}, \quad (13)$$

where $\overline{X_i}$ is the mean of X_i and $\overline{Y_i}$ is the mean of Y_i. The value of ρ lies between -1 and 1. If X_i and Y_i are completely correlated, ρ takes the value of 1 or -1; if X_i and Y_i are independent, ρ is zero. Thus, we compute the weight of $\mathcal{R}_i$ by summing up the absolute values of correlations between each feature within R_i and *C*, which is defined as:

$$W(\mathcal{R}_i) = \sum_{j=1}^{|\mathcal{R}_i|} \rho(F_j, C). \quad (14)$$

How to choose features from the selected representative sets. We adopt the idea of the p-value proposed in the alpha-investing algorithm [8, 33], to solve the selection of features from the selected representative sets. The alpha-investing algorithm uses linear regression to dynamically adjust the threshold of error reduction required for adding a new feature, as shown in Figure 5. To evaluate a feature whether it can be included by the predictive model so far, the alpha-investing algorithm defines three key parameters:

(1) The wealth w, represents the current acceptable number of future false positives.

(2) The threshold α, corresponds to the probability of including an irrelevant or redundant feature at each step. It is adjusted by the parameter w.

(3) The p-value is the probability that a feature coefficient would be judged to be non-zero when it is in fact zero. The p-value can be computed by linear regression.

The parameters $\Delta\alpha$ and w_0 are both user-adjustable parameter which can be selected to control the false discovery rate, and both of them are always set to 0.5 (False Discovery Rate: the number of features incorrectly included in the model divided by the total number of features included in the model). With those parameters, the alpha-investing algorithm sequentially considers each feature and dynamically adjusts the penalties *w* and α, for adding or discarding a feature upon its arrival [33]. More specially, a feature is added to the current model if its p-value is less than α. When a feature is added to the current model, the parameter *w* is increased while it is decreased when a feature is discarded in order to save enough wealth to add future features. The work of [9] has proved that the key success of alpha-investing is that it gives false discovery rate-style guarantees against overfitting in feature selection by dynamically adjusting the thresholds *w* and α on the p-value as a new feature to enter the model.

The Alpha-investing Algorithm

1: w_0 = 0.5; $\Delta\alpha$=0.5;
2: w=w_0; model = {};
3: i = 1; // index of features
4: Repeat
5: F_i = GetNewFeature(); // get a next feature
6: α= w/(2*i);
7: // Is the p-value of F_i below threshold?
8: if (Get-Pvalue(F_i, model) <=α) then accept F_i
9: model=model+F_i;
10: w= w+$\Delta\alpha$−α;
11: else // otherwise, reject F_i
12: w= w−α; // reduce wealth
13: endif
14: i = i+1;
15: Until no features are available
16: Return model.

Fig.5. The Alpha-investing Algorithm

But the alpha-investing algorithm cannot be directly applied to our problem, since a representative set is not a single feature, but a group of features. Motivated by the successes of the alpha-investing algorithm, we propose the SGAI algorithm (**S**election via **G**roup **A**lpha-**I**nvesting), to solve Eq.(7). The SGAI algorithm is summarized in Figure 6. With the representative sets, the SGAI algorithm consists of two Phases (Phases 3 and 4 in Figure 6). Phase 3 in Figure 6 is to calculate the predictive power of each representative set, while Phase 4 is to pick up one or a few feature(s) from the selected representative sets.

To access each representative set, at step 6, SGAI sets the normalized weight of each representative set $\mathcal{R}_i$ as its own initial wealth *w*, to measure how successful that a representative set has been generating predictive features. SGAI defines num_f(i) as the feature index of $\mathcal{R}_i$ to keep

track of the number of features that have been accessed in $\mathcal{R}_i$.

To select a representative set at each round, at the step 7, SGAI starts from the set with the highest weight of the set which has the highest probability of containing the useful features. As a selected representative set contributes one predictive feature to the current model, its corresponding wealth w will be increased, otherwise it will be decreased. For instance, if a representative set $\mathcal{R}_i$ contains a high fraction of predictive features, its wealth will soon become higher and features will be selected preferentially from this representative set. Meanwhile, if a representative set $\mathcal{R}_i$ has no predictive features, its wealth will soon be the lowest and it will be dropped entirely eventually.

The SGAI Algorithm

Input: F, C, Δα=0.5, and CSF={}
Phase 1: Identify representative sets $\mathcal{R}$
(1) $\Im$ =GetPC(C)
(2) K=|$\Im$| //Number of features in $\Im$
(3) For i=1 to K //Find $\mathcal{R}$ by Proposition 2
$\mathcal{R}_i = \text{GetPC}(F_i) \cup F_i, F_i \in \Im$
Endfor
Phase 2: Process $\mathcal{R}$ into K non-overlapping sets
(4) $\mathcal{R}_i \cap \mathcal{R}_j = \emptyset, i \neq j$
Phase 3: Assign the weights to representative sets
(5) For i=1 to K
if training data is discrete
$W(\mathcal{R}_i) = \sum_{j=1}^{|\mathcal{R}_i|} SU(C, F_j)$
else
$W(\mathcal{R}_i) = \sum_{j=1}^{|\mathcal{R}_i|} p(F_j, C)$
endif
Endfor
Phase 4: Select features from representative sets
(6) num_f(i)=1, $W(\mathcal{R}_i) = W(\mathcal{R}_i)/sum(W)$, i=1…$K$,
(7) Repeat
$\mathcal{R}_i = arg \max_{W(\mathcal{R}_i) \in W} W$
if $\mathcal{R}_i \neq$ null
X = GetNewFeature($\mathcal{R}_i$);
$\alpha = W(\mathcal{R}_i)/(2*$ num_f(i));
if (p-value(CFS, X) <=α)
$CFS = CFS \cup X$
$W(\mathcal{R}_i) = W(\mathcal{R}_i) + \Delta\alpha - \alpha$;
else
$W(\mathcal{R}_i) = W(\mathcal{R}_i) - \alpha$;
endif
num_f(i) = num_f(i) + 1;
endif
Until $\mathcal{R}_i == \emptyset$
(8) Return CFS.

Fig.6. The SGAI algorithm

CFS represents the currently selected feature set. With CFS and a currently selected $\mathcal{R}_i$, how does SGAI measure whether $F_i \in \mathcal{R}_i$ can be added to CFS or not? We use the p-value which is defined as the following.

$$p - value(CFS, F_i) = \exp\left(\frac{\left(\mathcal{L}(W\{\text{CFS} \cup \text{F}_i\}, \mathcal{R}, \text{C}) - \mathcal{L}(W(CFS), \mathcal{R}, C)\right)}{2\sigma^2}\right) \quad (15)$$

$\mathcal{L}(W\{CFS\}, \mathcal{R}, C)$ is the least square loss on the training set using the features in the *CFS*, $\mathcal{L}(W\{\text{CFS} \cup \text{F}_i\}, \mathcal{R}, C)$ is the loss after adding the feature F_i, and the variance is $\sigma^2 = \mathcal{L}(W(CFS), \mathcal{R}, C)/N$ where N is the number of the training data examples. For a new feature F_i, when the p-value of F_i is less than the threshold α, it is included into *CFS* (that is to say, the loss of $\mathcal{L}(W\{\text{CFS}\}, \mathcal{R}, C)$ is significantly reduced after adding F_i to *CFS*). Clearly, there is only one parameter, $\Delta\alpha$, for SGAI. As Alpha-investing does, it is set to 0.5 in the experiments. The parameter is not the key parameter and does not have a significant impact on SGAI.

Finally, we analyze the differences of our SGAI algorithm and the TIE* algorithm. TIE* discovers all Markov blankets for a data set customized for improving the causal induction mechanisms without missing causal variables, which is not for feature selection, whereas SGAI is specially customized for feature selection. Furthermore, at each iteration, TIE* preferentially finds a new Markov blanket on a new feature space by removing the previously discovered Markov blankets, and simply selects one feature from a set of correlated features (features in a seed set and its corresponding correlated features). SGAI attempts to choose a feature subset that maximizes the prediction power for classification from both a seed set and its corresponding correlated features simultaneously.

V. Experiments

A. Experiment Setup

We have chosen 16 benchmark data sets (Table 2). There are five data sets from the UCI machine learning repository (the first five data sets), three biomedical data sets (*hiva*, o*varian-cancer*, and *breast-cancer*), four NIPS 2003 feature selection challenge data sets (*arcene*, *dexter*, *dorothea*, and *madelon*), and four public microarray data sets (the last four data sets) [28]. Those 16 data sets cover a wide range of real-world application domains, including clinical images, gene expressions, ecology, text categorization, and molecular biology. The dimensionality of the 16 data sets (from 22 to 100,000) and sample sizes (from sixty to thousands) represent practical applications.

We use two classifiers, the Naïve Bayes (NB) classifier provided by the Matlab statistical toolbox and the SVM classifier provided by the LIBSVM library[1]. For those data sets, we use 10-fold cross-validation, and report prediction errors and their corresponding standard deviations of NB and SVM in our experiments.

Our comparative study uses four state-of-the-art Markov blanket filters, including IAMB [23], MMMB [24], PCMB [18], and HITON-MB [3], the state-of-the-art multiple Markov blanket discovery algorithm TIE* [21], the recent SRS algorithm [31], and four well-established feature selection algorithms, FCBF [29], mRMR [19], SPSF-LAR [32], and MRF [6]. The parameters in our experiments are set as follows.

(1) We use nested N-fold cross-validation [22] on each

[1] The LIBSVM library is available at www.csie.ntu.edu.tw/~cjlin/libsvm/

training data set to get descent λ_1 and λ_2 for SRS. The inner loop is used to try different parameters for feature selection and classification while the outer loop tests the best configuration on an independent test data set.

(2) The significant level is set to 0.01 for IAMB, MMMB, PCMB, and HITON-MB.

(3) IAMB, MMMB, PCMB, HITON-PC and HITON-MB deal with discrete data with the G^2-test while IAMB, HITON-PC and HITON-MB handle continuous data with the Fisher's z-test[2].

(4) To validate whether SGAI and its rivals have no significant difference in prediction errors, we conduct the Friedman test at a 95% significance level [7], under the null-hypothesis, which states that the performance of SGAI and that of its rivals have no significant difference, and calculate the average ranks using the Friedman test (for calculating the average ranks, please see [7]). When the null-hypothesis at the Friedman test is rejected, we proceed with the Nemenyi test [7] as a post-hoc test. With the Nemenyi test, the performance of two methods is significantly different if the corresponding average ranks differ by at least the critical difference (for calculating the critical difference, please see [7]).

(5) All experiments are conducted on a computer with Inter(R) i7-2600 3.4GHz CPU and 12GB memory.

TABLE 2 SUMMARY OF THE BENCHMARK DATA SETS.
(#F: NUMBER OF FEATURES, #I: NUMBER OF INSTANCES)

Dataset	#F	#I	Dataset	#F	#I
spect	22	267	madelon	500	2,000
wdbc	30	569	colon	2,000	62
spectf	44	267	prostate	6,033	102
promoter	57	106	leukemia	7,129	72
infant	86	5,337	lung-cancer	12,533	181
arcene	10,000	100	breast-cancer	17,816	286
dexter	20,000	300	ovarian-cancer	2,190	216
dorothea	100,000	800	hiva	1,617	4,229

B. Comparison of SGAI with SRS and RES

In this section, in addition to compare SGAI with SRS using the NB and SVM classifiers respectively, we further investigate the set of the union of representative sets. RES (***RE***presentative ***S***ets) means that we use the union of representative sets as a feature subset and calculate its classification errors.

1) Comparison on Prediction Errors

Tables 3 and 4 summarize the prediction errors and the standard deviations of SGAI against SRS and RES using NB and SVM, respectively. We conduct paired t-tests at a 95% significance level and summarize the win/tie/lose counts of SGAI against SRS and RES in the last rows of Tables 3 and 4. The lowest errors are highlighted in bold face.

With the counts of win/tie/loss in Table 3, we observe that SGAI outperforms both SRS and RES using the NB classifier. To further evaluate the performance of SGAI against SRS and RES, we conduct the Friedman test at a 95% significance level under the null-hypothesis, which states that whether the performance of SGAI and that of SRS and RES have no significant difference in prediction errors. The null-hypothesis is rejected, and the average ranks for SGAI, SRS, and RES are 1.22, 2.13, and 2.66, respectively (the lower the average rank, the better the performance in prediction errors).

TABLE 3 PERFORMANCE OF SGAI, SRS, AND RES USING NB
(A/B: A DENOTES PREDICTION ERROR AND B IS STANDARD DEVIATION)

Dataset	SGAI	SRS	RES
spect	**0.2918/0.0916**	0.3106/0.0896	0.3106/0.0896
wdbc	**0.0562/0.0245**	0.0756/0.0407	0.0667/0.0394
spectf	**0.1607/0.0738**	0.2097/0.0746	0.2508/0.0741
promoter	**0.0336/0.0831**	0.0340/0.1554	0.0973/0.0944
infant	**0.0449/0.0080**	0.0545/0.0090	0.0618/0.0090
arcene	**0.3100/0.1449**	0.3300/0.1767	0.3600/0.1897
dexter	**0.0967/0.0483**	0.2000/0.0968	0.1000/0.0785
dorothea	0.0600/0.0269	**0.0488/0.0246**	0.0575/0.0329
madelon	**0.3590/0.0464**	0.3700/0.0418	0.3705/0.0442
colon	**0.0929/0.1049**	0.1262/0.1235	0.2714/0.1472
prostate	**0.0482/0.0509**	0.0582/0.0689	0.0873/0.0708
leukemia	**0.0125/0.0395**	**0.0125/0.0395**	0.0286/0.0602
lung-cancer	**0.0056/0.0176**	0.0442/0.0438	**0.0111/0.0234**
breast-cancer	**0.0911/0.0384**	0.1223/0.0600	0.1222/0.0600
ovarian-cancer	0.0883/0.0467	**0.0837/0.0615**	0.1208/0.0728
hiva	**0.0470/0.0080**	0.0603/0.0090	0.0674/0.0111
average rank	**1.22**	2.13	2.66
win/tie/loss	-	11/4/1	13/3/0

TABLE 4 PERFORMANCE OF SGAI, SRS, AND RES USING SVM
(A/B: A DENOTES PREDICTION ERROR AND B IS STANDARD DEVIATION)

Dataset	SGAI	SRS	RES
spect	**0.2997/0.0615**	**0.2997/0.0615**	**0.2997/0.0615**
wdbc	0.0281/0.0220	0.0580/0.0397	**0.0211/0.0199**
spectf	**0.1235/0.0580**	0.1799/0.0391	0.2023/0.0399
promoter	0.2472/0.1448	0.3400/0.1554	0.2563/0.1296
infant	**0.0431/0.0070**	0.0446/0.0056	**0.0431/0.0070**
arcene	**0.1900/0.1101**	**0.1900/0.1595**	0.2000/0.1633
dexter	**0.1067/0.0410**	0.1167/0.0451	0.1400/0.0604
dorothea	0.0600/0.0275	**0.0550/0.0206**	0.0588/0.0260
madelon	0.1885/0.0225	0.1765/0.0226	**0.1635/0.0268**
colon	0.1238/0.1554	**0.1214/0.1689**	0.1738/0.1361
prostate	0.0582/0.0502	**0.0382/0.0494**	0.0673/0.0789
leukemia	**0.0125/0.0395**	0.0411/0.0663	0.0411/0.0945
lung-cancer	**0.0111/0.0234**	**0.0111/0.0246**	**0.0111/0.0234**
breast-cancer	**0.0911/0.0448**	0.1047/0.0618	0.1188/0.0576
ovarian-cancer	0.0697/0.0515	**0.0554/0.0472**	0.0974/0.0467
hiva	0.0354/0.0017	**0.0351/0.0013**	0.0351/0.0013
average rank	**1.81**	1.91	2.28
win/tie/loss	-	6/7/3	7/8/1

Then we proceed with the Nemenyi test as a post-hoc test. With the Nemenyi test, the performance of two methods is significantly different if the corresponding average ranks differ by at least the critical difference. With the Nemenyi test, the critical difference is up to 0.83. Thus, we can observe that SGAI is significantly better than both SRS and RES using the NB classifier in the prediction errors.

Table 4 gives the results of SGAI, SRS, and RES using SVM. We can see that SGAI also outperforms both SRS and RES using the SVM classifier. To further evaluate the performance of SGAI against SRS and RES, we conduct the Friedman test. The null-hypothesis is accepted, and the

[2] In our preliminary version [31], IAMB, MMMB, PCMB, HITON-PC, and HITON-MB are all conducted on discrete or discretized data.

average ranks for SGAI, SRS, and RES are 1. 81, 1.91, and 2.28, respectively.

2) Comparison on Numbers of Selected Features

Figures 7 and 8 give the result of the numbers of selected features of SGAI and SRS. SGAI selects fewer features than SRS on most data sets.

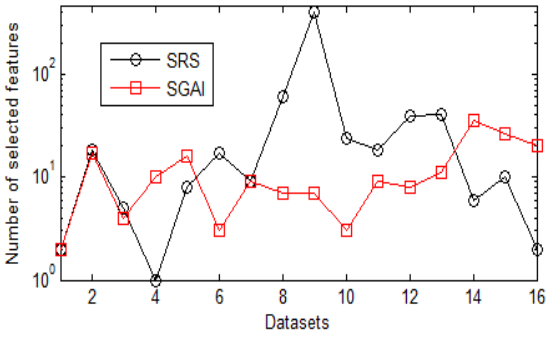

Fig. 7. Number of selected features of SGAI and SRS (the labels of the x-axis from 1 to 16 denote the data sets: 1. spect, 2. wdbc, 3. spectf, 4. promoter, 5. infant, 6. arcene, 7. dexter, 8. dorothea, 9. madelon, 10. colon, 11. prostate, 12.leukemia, 13. lung-cancer, 14. breast-cancer, 15. ovarian-cancer, and 16. hiva)

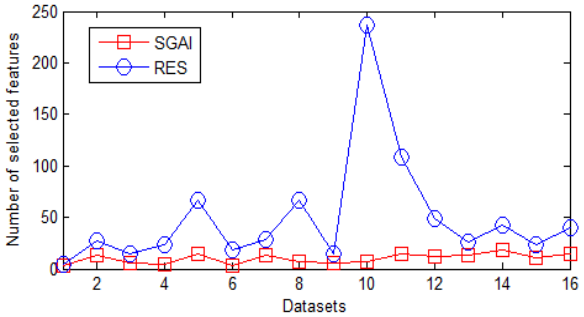

Fig.8. Number of selected features of SGAI against RES (the labels of the x-axis from 1 to 16 are the same as those in Figure 7.)

More importantly, SGAI does not require the user-defined parameters λ_1 and λ_2 in SRS while it is a nontrivial problem for SRS to get a set of descent λ_1 and λ_2. Meanwhile, there is only one parameter, $\Delta\alpha$, for SGAI. It is not a key parameter of SGAI and does not have a significant impact on SGAI. In Figure 8, SGAI selects fewer features than RES.

As for the running time, given the representative sets and the decent input parameters λ_1 and λ_2, SGAI has almost the same running time as SRS. Clearly, the computational cost of the SGAI algorithm mainly spends on calculating representative sets. Thus, we don't report running time of SRS, RES, and SGAI here.

*C. Comparison with TIE**

In this section, we compare our SGAI algorithm with the state-of-the-art multiple Markov blanket discovery algorithm, TIE*, which attempts to find all Markov blankets in real data in non-faithful conditions for improving causal induction by avoiding missing causative variables.

In our experiments, TIE* is parameterized with HITON-PC as the base Markov blanket induction algorithm in order to be consistent with our SGAI algorithm which also employs HITON-PC to discover representative sets (in our preliminary version, TIE* employed a semi-interleaved HITON-PC), and a classification error as a criterion that verifies whether a new feature subset is a Markov blanket of a class attribute. With the same parameter setting of the TIE* algorithm in [21], the parameter alpha of HITON-PC is set to 0.05. We select the Markov blanket with the lowest prediction error from all of the Markov blankets discovered by TIE*.

In the following tables and figures, we report the comparison results of SGAI and TIE*. The highest prediction errors are highlighted in bold face. We select the lowest prediction errors of the *dexter* and *breast-cancer* data sets within 3 days for TIE* due to long running time (exceeding three days).

TABLE 5 PERFORMANCE OF SGAI AND TIE* USING NB
(A/B: A DENOTES PREDICTION ERROR AND B IS STANDARD DEVIATION)

Dataset	SGAI	TIE*
spect	**0.2918/0.0916**	0.3108/0.0530
wdbc	0.0562/0.0245	**0.0369/0.0280**
spectf	**0.1607/0.0738**	0.1981/0.0375
promoter	**0.0336/0.0831**	0.0491/0.0701
infant	**0.0449/0.0080**	0.0467/0.0094
arcene	0.3100/0.1449	**0.1700/0.1251**
dexter	**0.0967/0.0483**	0.1600/0.0858
dorothea	**0.0600/0.0269**	**0.0600/0.0269**
madelon	**0.3590/0.0464**	0.3795/0.0471
colon	**0.0929/0.1049**	0.1381/0.1894
prostate	**0.0482/0.0509**	0.0682/0.0667
leukemia	**0.0125/0.0395**	**0.0125/0.0395**
lung-cancer	**0.0056/0.0176**	0.0091/0.0324
breast-cancer	0.0911/0.0384	**0.0732/0.0497**
ovarian-cancer	0.0883/0.0467	**0.0604/0.0667**
hiva	0.0470/0.0080	**0.030/ 0.0049**
win/tie/loss	-	8/4/4
average ranks	**1.38**	1.63

TABLE 6 PERFORMANCE OF SGAI AND TIE* USING SVM
(A/B: A DENOTES PREDICTION ERROR AND B IS STANDARD DEVIATION)

Dataset	SGAI	TIE*
spect	**0.2997/0.0615**	**0.2997/0.0615**
wdbc	**0.0281/0.0220**	0.0334/0.0292
spectf	**0.1235/0.0580**	0.2058/0.0175
promoter	**0.2472/0.1448**	0.2482/0.1538
infant	**0.0431/0.0070**	0.0435/0.0061
arcene	**0.1900/0.1101**	**0.1900/0.0994**
dexter	**0.1067/0.0410**	0.1467/0.1485
dorothea	**0.0600/0.0275**	0.0613/0.0303
madelon	**0.1885/0.0225**	0.3795/0.0515
colon	0.1238/0.1554	**0.0905/0.1610**
prostate	**0.0582/0.0502**	0.0782/0.0915
leukemia	**0.0125/0.0395**	0.0554/0.0983
lung-cancer	**0.0111/0.0234**	**0.0111/0.0421**
breast-cancer	0.0911/0.0448	**0.0800/0.0511**
ovarian-cancer	**0.0697/0.0515**	0.0740/0.0320
hiva	0.0354/0.0017	**0.0333/0.0038**
win/tie/loss	-	5/9/2
average ranks	**1.28**	1.72

Using NB, in Table 5, with the counts of win/tie/loss, we observe SGAI only loses four times against TIE*. With the Friedman test, the null-hypothesis is not rejected, and the average ranks for SGAI and TIE* are 1.38 and 1.63, respectively. Using SVM, in Table 6, by the counts of win/tie/loss, SGAI only loses twice against TIE*. With the Friedman test, the null-hypothesis is rejected, and the average

ranks for SGAI and TIE* are 1.28 and 1.72, respectively. Then we proceed with the Nemenyi test as a post-hoc test, and the critical difference is up to 0.49. Thus, we can conclude that SGAI is significantly better than TIE* using the SVM classifier in the prediction errors.

Accordingly, using both NB and SVM, SGAI achieves highly competitive performance against TIE* without an exhaustive search over all candidate Markov blankets.

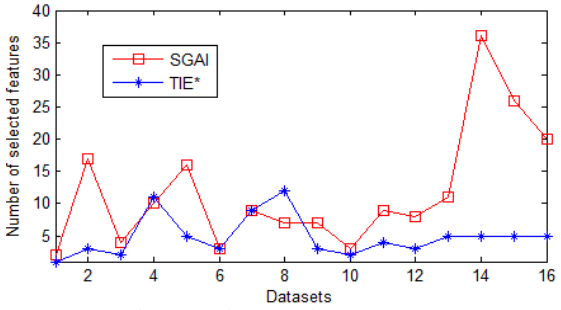

Fig.9. Numbers of selected features of SGAI and TIE* (the labels of the x-axis from 1 to 16 are the same as those in Figure 7.)

TABLE 7 RUNNING TIME (IN SECONDS)

Dataset	SGAI	TIE*
spect	1	1
wdbc	2	1
spectf	2	1
promoter	1	143
infant	16	9
arcene	16	13
dexter	27	>3 days
dorothea	309	6,665
madelon	2	2
colon	1,039	2
prostate	27	13
leukemia	24	159
lung-cancer	41	76
breast-cancer	61	>3 days
ovarian-cancer	6	7
hiva	17	30

Why is SGAI not worse than TIE*? A possible explanation is that since TIE* finds a new Markov blanket on a new feature space by removing the previously discovered Markov blankets from the data set at each round, TIE* simply selects one feature from a set of correlated features, whereas SGAI considers both strongly relevant features and their corresponding correlated features simultaneously, and this might be beneficial to reduce classification error. This also explains why SGAI selects more features than TIE* as shown in Figure 9. For example, on the *infant* dataset, SGAI gets 13 representative sets, and features in each group are strongly correlated. SGAI further selects six representative sets and picks up more than two features from those groups respectively while the Markov blankets selected by TIE* only contain 5 strongly relevant features which attain the lowest errors among all Markov blankets.

From Table 7, we can see that TIE* fails on the *dexter* and *breast-cancer* data sets due to long running time (exceeding three days). But on the *colon* and *leukemia* data sets, TIE* is much faster than SGAI. A possible explanation is that in those data sets, there are a few Markov blankets for TIE* while there are many correlated features corresponding to features in the seed set to be dealt with by SGAI.

In summary, instead of an exhaustive search for all Markov blankets, SGAI minimizes the prediction errors from representative sets with reasonable running time.

D. Comparison with Other Markov Blanket Filters

Tables 9 and 10 summarize the prediction errors of SGAI against the state-of-the-art Markov blanket filters, HITON-MB, IAMB, PCMB, and MMMB. The highest prediction errors are highlighted in bold face. Note that some entries in Tables 9 and 10 are marked by "-", which means that the method fails to return any result within a reasonable response time (i.e., 72 hours for a single training in our case). For example, MMMB fails on the *dorothea* data set, due to long running time. With the win/tie/lose counts of SGAI against HITON-MB, IAMB, PCMB and MMMB in the last row of Table 9, we observe that SGAI is superior to HITON-MB, IAMB, PCMB and MMMB on most data sets.

Meanwhile, using the Friedman test at a 95% significance level, the null-hypothesis is rejected, and the average ranks for SGAI, HITON-MB, IAMB, PCMB and MMMB are 1.78, 2.88, 3.47, and 1.88, respectively.

Then we proceed with the Nemenyi test as a post-hoc test. With the Nemenyi test, the critical difference is up to 1.17. Thus, we can conclude that SGAI is significantly better than PCMB using the NB classifier in the prediction errors. Since MMMB fails on the *dorothea* data set, we don't compare SGAI with MMMB using the Friedman test.

According to Table 10 using SVM, we can see that SGAI also outperforms HITON-MB, IAMB, PCMB and MMMB on most data sets. Furthermore, SGAI is never worse than PCMB in prediction errors. With the Friedman test at a 95% significance level, the average ranks for SGAI, HITON-MB, IAMB, PCMB and MMMB are 1.97, 2.34, 2.97, and 2.72, respectively. As for the number of selected features, in Figure 10, SGAI is also very competitive with its rivals, although it considers not only the seed set of features but also their corresponding correlated features. From Table 8, on running time (in seconds), SGAI is also very competitive with the other MB filters (excluding the *dexter*, *madelon*, *lung-cancer* and *breast-cancer* data sets), even though SGAI needs to consider not only the seed set of features but also their corresponding correlated features. Since the PCMB algorithm is implemented in the C language, we don't give its running time here. We also don't present the *dorothea* and *leukemia* data sets as MMMB fails on those two data sets.

In summary, our empirical study has demonstrated that when we treat data under non-faithful conditions, Markov blankets selected by the existing single Markov blanket feature selection methods may not be a best feature subset that minimizes the prediction error for classification. Our SGAI algorithm may effectively and efficiently handle Markov blanket feature selection in real data with representative sets. More importantly, with representative sets, SGAI could efficiently find a best feature subset for feature selection without an exhaustive search over an unknown space of all Markov blankets in a real data set.

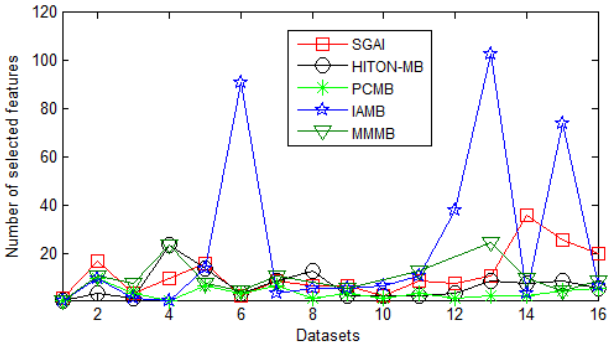

Fig.10. Numbers of selected features of SGAI vs. other Markov Blanket Filters (the labels of the x-axis are the same as the labels in Fig.7).

TABLE 8 RUNNING TIME (IN SECONDS)

Dataset	SGAI	IAMB	MMMB	HITON-MB
spect	1	0.1	0.1	0.1
wdbc	2	1	2	1
spectf	2	0.1	1	1
promoter	1	0.1	1	1
infant	16	1	1	2
arcene	16	367	16	16
dexter	27	19	36	37
dorothea	309	248	/	305
madelon	2	1	2	1
colon	1,039	2	7	2
prostate	27	325	89	12
leukemia	24	208	23	13
lung-cancer	41	1,336	39	47
breast-cancer	61	6,085	63	50
ovarian-cancer	6	66	6	8
hiva	17	8	43	18

E. Comparison with Other Feature Selectors

Tables 11 and 12 summarize the prediction errors of SGAI against two well-established feature selection methods, FCBF and mRMR, and two state-of-the-art feature selection algorithms, SPSF-LAR and MRF. The highest prediction errors are highlighted in bold face.

Since SGAI selects no more than 60 features to get the lowest prediction error on all 16 data sets, for SPSF-LAR, MRF, and mRMR, we use the selected feature subset whose size ranges from 1 to 15 for five UCI data sets and choose the lowest prediction errors achieved by NB and SVM, while for the remaining 11 high-dimensional data sets, we use the top 5, 10, 15, ..., 60 features selected by each algorithm.

We conduct paired t-tests at a 95% significance level and summarize the win/tie/lose counts of SGAI against its four rivals in the last rows of Tables 11 and 12. Meanwhile, with the Friedman test at a 95% significance level, the average ranks calculated from the Friedman test for SGAI, SPSF-LAR, MRF, FCBF and mRMR are 2.78, 2.88, 3.78, 2.5, and 3.06 respectively in Table 11 using NB, while using SVM, the average ranks are 2.72, 2.56, 2.87,3.84, and 3.00, respectively, in Table 12. Accordingly, with the counts of win/tie/loss and the average ranks in Tables 11 and 12, we can conclude that SGAI achieves highly competitive performance against the four well-established feature selection algorithms.

VI. CONCLUSION

In this paper, we explored Markov blanket feature selection by assuming that a probability distribution may not be faithfully represented by a Bayesian network. To tackle this issue, we extended the concept of Markov blankets and proposed the concept of representative sets. With representative sets, we presented the SGAI algorithm for Markov blanket feature selection under the non-faithful condition. The experimental results have shown that the SGAI algorithm outperforms both state-of-the-art Markov blanket feature selectors and other well-established feature selection methods using real-world data.

ACKNOWLEDGMENTS

A shorter, preliminary version of this paper with the title "Markov Blanket Feature Selection with Non-Faithful Data Distributions" was published in the Proceedings of the 13th IEEE International Conference on Data Mining (ICDM'13), pp. 857-866. This work is partly supported by the National 973 Program of China (under grant 2013CB329604), the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education, China (under grant IRT13059), and the National Natural Science Foundation of China (under grants 61229301, 61372191, and 61572492).

REFERENCES

[1] A. Farahat , A. Elgohary, A. Ghodsi, and M. Kamel. (2015) Greedy column subset selection for large-scale data sets. Knowledge and Information Systems, 45(1), 1-34.

[2] F. Ahmed K., K. Benabdeslem, and N. Tale. (2015) Soft-constrained Laplacian score for semi-supervised multi-label feature selection. Knowledge and Information Systems, 1-24.

[3] 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.

[4] 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.

[5] R. Chakraborty and N. R. Pal. (2015) Feature selection using a neural framework with controlled redundancy. IEEE Transactions on Neural Networks and Learning Systems, 26(1), 35-50.

[6] 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.

[7] J. Demšar. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1-30.

[8] P. S. Dhillon, D. Foster and L. Ungar. (2010) Feature selection using multiple streams. AISTATS'10, 153-160.

[9] D. P. Foster and R. A. Stine. (2008) α-investing: A procedure for sequential control of expected false discoveries. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 70(2):429–444.

[10] D. Koller and M. Sahami. (1996) Toward Optimal Feature Selection.

ICML'96, 284-292.
[11] R. Kohavi and G. H. John. (1997) Wrappers for Feature Subset Selection. Artificial Intelligence, 97, 273-324.
[12] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, and X. Zhou. (2015) Semisupervised Feature Selection via Spline Regression for Video Semantic Recognition. IEEE Transactions on Neural Networks and Learning Systems, 26(2), 252-264.
[13] Y. Li, J. Si, G. Zhou, S. Huang, and S. Chen. (2015) FREL: A Stable Feature Selection Algorithm. IEEE Transactions on Neural Networks and Learning Systems, 26(7), 1388-1402.
[14] H. Liu and L. Yu. (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering, 17(4), 491-502.
[15] X. Liu, L. Wang, J. Zhang, J. Yin, and H. Liu. (2014) Global and local structure preservation for feature selection. IEEE Transactions on Neural Networks and Learning Systems, 25(6), 1083-1095.
[16] D. Margaritis and S. Thrun. (2000) Bayesian Network Induction via Local Neighborhoods. In Advances in Neural Information Processing Systems 1999. Denver, Colorado, USA: The MIT Press.
[17] J. Pearl. (1988) Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, CA.
[18] 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.
[19] 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.
[20] J. R. Quinlan. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
[21] A. Statnikov, N. Lytkin, J. Lemeire and F. C. Aliferis. (2013) Algorithms for Discovery of Multiple Markov Boundaries. Journal of Machine Learning Research 14, 499-566.
[22] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin and S. Levy. (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics, 21(5), 631-643.
[23] I. Tsamardinos and C. F. Aliferis. (2003) Towards Principled Feature Selection: Relevancy, Filters and Wrappers. AI & Statistics'03.
[24] 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.
[25] D. Wang, W. Ding, K. Yu, X. Wu, P. Chen, D. L. Small, and S. Islam. (2013) Towards long lead forecasting of extreme flood events: a data mining framework for precipitation cluster precursors identification. In KDD'13, 1285-1293.
[26] X. Wu, K. Yu, W. Ding, H. Wang, and X. Zhu. (2013) Online Feature Selection with Streaming Features. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(5): 1178-1192 (2013).
[27] J. Xiao, Y. Xiao, A. Huang, D. Liu, and S. Wang. (2015) Feature-selection-based dynamic transfer ensemble model for customer churn prediction. Knowledge and Information Systems, 43(1), 29-51.
[28] L. Yu, C. Ding, and S. Loscalzo. (2008) Stable Feature Selection via Dense Feature Groups. KDD'08, 803-811.
[29] L. Yu and H. Liu. (2004) Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research, 5: 1205-1224.
[30] 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. 25(12): 2721-2739.
[31] K. Yu, X. Wu, Z. Zhang, Y. Mu, H. Wang, and W. Ding. (2013) Markov Blanket Feature Selection with Non-faithful Data Distributions. IEEE ICDM'13, 857-866.
[32] 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.
[33] J. Zhou, D. Foster, R.A. Stine and L.H. Ungar. (2006) Streamwise feature selection. Journal of Machine Learning Research, 7:1861-1885.

Kui Yu received his Ph.D. degree in Computer Science in 2013 from the Hefei University of Technology, China. From 2013 to 2015, he was a postdoctoral fellow in the School of Computing Science, Simon Fraser University, Canada. He is currently a research fellow in the School of Information Technology and Mathematical Sciences, University of South Australia, Australia. His research interests include causal discovery and feature selection.

Xindong Wu received the bachelor's and master's degrees in computer science from the Hefei University of Technology, China, and the PhD degree in artificial intelligence from the University of Edinburgh, Edinburgh, United Kingdom.

He is currently Director and Professor in the School of Computing and Informatics at the University of Louisiana, Lafayette, USA. His research interests include data mining, Big Data analytics, knowledge-based systems, and Web information exploration. He is the steering committee chair of the IEEE International Conference on Data Mining, the editor-in chief of Knowledge and Information Systems (KAIS, by Springer), and a series editor-in-chief of the Springer Book Series on Advanced Information and Knowledge Processing (AI&KP). He was the editor-in-chief of the IEEE Transactions on Knowledge and Data Engineering (TKDE) between 2005 and 2008. He served as program committee chair/co-chair for ICDM '03 (the 2003 IEEE International Conference on Data Mining), KDD-07 (the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining), and CIKM 2010 (the 19th ACM Conference on Information and Knowledge Management). He is a Fellow of the IEEE and AAAS.

Wei Ding received her Ph.D. degree in Computer Science from the University of Houston in 2008. She is an Associate Professor of Computer Science in the University of Massachusetts Boston. Her research interests include data mining, machine learning, artificial intelligence, computational semantics, and with applications to astronomy, geosciences, and environmental sciences. She has published more than 105

referred research papers, 1 book, and has 2 patents. She is an Associate Editor of Knowledge and Information Systems (KAIS) and an editorial board member of the Journal of Information System Education (JISE), the Journal of Big Data, and the Social Network Analysis and Mining Journal. She is the recipient of a Best Paper Award at the 2011 IEEE International Conference on Tools with Artificial Intelligence (ICTAI), a Best Paper Award at the 2010 IEEE International Conference on Cognitive Informatics (ICCI), a Best Poster Presentation award at the 2008 ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPAITAL GIS), and a Best PhD Work Award between 2007 and 2010 from the University of Houston. Her research projects are sponsored by NIH, NASA, and DOE. She is an IEEE senior member and an ACM senior member.

Yang Mu received his B.S. and Ph.D. degrees from Jilin University and the University of Massachusetts Boston in 2008 and 2015 respectively. Prior to his PhD study, he worked at Nanyang Technological University as a research assistant. He is currently a research scientist in Facebook Inc. His research interests include machine learning and data mining.

Hao Wang received his B.S. degree from the Department of Electrical Engineering and Automation at the Shanghai Jiao Tong University, China, and his M.S. and Ph.D. degrees in Computer Science from the Hefei University of Technology. His interests are in Artificial Intelligence and Robotics and Knowledge Engineering. He is a Professor of the School of Computer Science and Information Engineering, Hefei University of Technology, China.

TABLE 9 RESULTS OF EACH ALGORITHM USING NB
(A/B: A DENOTES PREDICTION ERROR WHILE B REPRESENTS STANDARD DEVIATION)

Dataset	SGAI	IAMB	PCMB	HITON-MB	MMMB
spect	**0.2918/0.0916**	0.3108/0.0529	0.3108/0.0529	0.2997/0.0615	0.3108/0.0529
wdbc	0.0562/0.0245	0.0456/0.0343	0.0685/0.0291	**0.0387/0.0318**	0.0737/0.0306
spectf	**0.1607/0.0738**	0.1907/0.0555	0.2165/0.0762	0.2392/0.0618	0.2427/0.0733
promoter	**0.0336/0.0831**	0.2000/0.086	0.4146/0.0955	0.0972/0.0944	0.0972/0.0944
infant	**0.0449/0.0080**	0.0474/0.009	0.0603/0.0038	0.0487/0.0100	0.0504/0.0095
arcene	0.3100/0.1449	0.3100/0.1524	0.2800/0.1135	**0.17/0.1159**	0.3400/0.1174
dexter	**0.0967/0.0483**	0.2500/0.0671	0.2567/0.0630	0.24/0.062	0.2533/0.0706
dorothea	**0.0600/0.0269**	0.0663/0.0167	0.0988/0.0040	0.0613/0.0208	/
madelon	**0.3590/0.0464**	0.3740/0.0367	0.4315/0.0274	0.37/0.0418	0.3700/0.0418
colon	**0.0929/0.1049**	0.1571/0.1616	0.1405/0.1495	0.1381/0.1894	0.2857/0.1694
prostate	**0.0482/0.0509**	0.2418/0.1308	0.0764/0.0729	0.0682/0.0667	0.1164/0.1003
leukemia	**0.0125/0.0395**	0.1629/0.1420	0.0554/0.0716	0.0268/0.0566	0.0554/0.0716
lung-cancer	**0.0056/0.0176**	0.0500/0.0484	0.0664/0.0439	0.0167/0.0268	0.0222/0.0388
breast-cancer	**0.0911/0.0384**	0.1297/0.0585	0.1156/0.0376	0.108/0.0466	0.0943/0.0401
ovarian-cancer	0.0883/0.0467	0.2506/0.0751	0.3996/0.1328	0.1026/0.0703	**0.0788/0.0787**
hiva	0.0470/0.0080	**0.0312/0.0040**	0.0317/0.0042	0.0315/0.0039	0.0317/0.0038
average ranks	**1.78**	2.88	3.47	1.88	-
win/tie/loss	-	10/3/3	13/1/2	8/4/4	10/3/2

TABLE 10 RESULTS OF EACH ALGORITHM USING SVM
(A/B: A DENOTES PREDICTION ERROR WHILE B REPRESENTS STANDARD DEVIATION)

Dataset	SGAI	IAMB	PCMB	HITON-MB	MMMB
spect	**0.2997/0.0615**	**0.2997/0.0615**	**0.2997/0.0615**	0.3071/0.0670	**0.2997/0.0615**
wdbc	0.0281/0.0220	**0.0263/0.0289**	0.0316/0.0199	0.0369/0.0267	0.0386/0.0318
spectf	**0.1235/0.0580**	0.2097/0.0184	0.2058/0.0175	0.2058/0.0175	0.2095/0.0297
promoter	0.2472/0.1448	**0.2182/0.0793**	0.3945/0.1074	0.2564/0.1296	0.2564/0.1296
infant	**0.0431/0.0070**	0.0438/0.0065	0.0582/0.0490	0.0438/0.0075	0.0438/0.0057
arcene	**0.1900/0.1101**	0.4600/0.0700	**0.1900/0.1920**	0.2000/0.1054	0.2100/0.1449
dexter	**0.1067/0.0410**	0.1800/0.0549	0.2900/0.1899	0.5000/0	0.3867/0.1841
dorothea	0.0600/0.0275	**0.0525/0.0317**	0.0975/0.0052	0.0600/0.0305	/
madelon	0.1885/0.0225	0.3795/0.0404	0.3440/0.0431	**0.1625/0.0226**	**0.1625/0.0226**
colon	0.1238/0.1554	**0.0786/0.1328**	0.1405/0.1689	0.1095/0.1286	0.2548/0.1684
prostate	0.0582/0.0502	**0.0564/0.0887**	0.0582/0.0502	0.0782/0.0915	0.1055/0.0920
leukemia	**0.0125/0.0395**	0.0393/0.0966	0.0411/0.0663	0.0286/0.0904	0.0411/0.0663
lung-cancer	0.0111/0.0234	0.0167/0.0268	0.0111/0.0234	0.0111/0.0234	**0.0056/0.0176**
breast-cancer	**0.0911/0.0448**	0.1333/0.0606	0.1121/0.0522	0.1220/0.0432	0.1015/0.0350
ovarian-cancer	0.0697/0.0515	**0.0231/0.0331**	0.3933/0.0789	0.0416/0.0404	0.0465/0.0444
hiva	0.0354/0.0017	**0.0333/0.0038**	**0.0333/0.0038**	**0.0333/0.0038**	**0.0333/0.0038**
average ranks	**1.97**	2.34	2.97	2.72	-
win/tie/loss	-	6/7/3	10/6/0	6/6/4	7/6/2

TABLE 11 RESULTS OF SGAI, SPSF-LAR, MRF, FCBF, AND MRMR USING NB
(A/B: A DENOTES PREDICTION ERROR WHILE B REPRESENTS STANDARD DEVIATION)

Dataset	**SGAI**	**SPSF-LAR**	**MRF**	**FCBF**	**mRMR**
spect	0.2918/0.0916	0.3144/0.0962	0.3034/0.0624	**0.2698/0.0790**	0.3144/0.0962
wdbc	**0.0562/0.0245**	0.0587/0.0315	0.0510/0.0303	0.0580/0.0321	0.0587/0.0315
spectf	**0.1607/0.0738**	0.2585/0.0689	0.2053/0.0940	0.2129/0.0690	0.2585/0.0689
promoter	**0.0336/0.0831**	0.1181/0.0922	0.0564/0.0486	0.1005/0.0937	0.1181/0.0922
infant	**0.0449/0.0080**	0.0545/0.0096	0.0463/0.0076	0.0467/0.0096	0.0545/0.0096
arcene	0.3100/0.1449	**0.2163/0.1317**	0.3000/0.1826	0.2800/0.1135	**0.2163/0.1317**
dexter	**0.0967/0.0483**	0.2637/0.0912	0.1367/0.0637	0.2267/0.0617	0.2637/0.0912
dorothea	0.0600/0.0269	0.0638/0.0291	**0.0163/0.0132**	0.0538/0.0240	0.0638/0.0291
madelon	0.3590/0.0464	0.3661/0.0361	0.3705/0.0317	0.3482/0.0361	0.3661/0.0361
colon	**0.0929/0.1049**	0.1240/0.1401	0.1428/0.1156	0.1244/0.1165	0.1240/0.1401
prostate	**0.0482/0.0509**	0.0653/0.0689	0.0576/0.0654	0.0650/0.0717	0.0653/0.0689
leukemia	**0.0125/0.0395**	0.0611/0.0998	0.0286/0.0602	0.0517/0.0926	0.0611/0.0998
lung-cancer	**0.0056/0.0176**	0.0400/0.0456	0.0719/0.0590	0.0506/0.0401	0.0400/0.0456
breast-cancer	0.0911/0.0384	0.0908/0.0547	**0.0837/0.0439**	0.0955/0.0548	0.0908/0.0547
ovarian-cancer	0.0883/0.0467	0.1636/0.0833	**0.0788/0.0440**	0.1092/0.0627	0.1636/0.0833
hiva	0.0470/0.0080	0.0351/0.0013	**0.0304/0.0040**	0.0361/0.0013	0.0351/0.0013
average ranks	2.78	2.88	3.78	**2.50**	3.06
win/tie/loss	-	6/4/6	8/4/4	5/6/5	7/4/5

TABLE 12 RESULTS OF SGAI, SPSF-LAR, MRF, FCBF, AND MRMR USING SVM
(A/B: A DENOTES PREDICTION ERROR WHILE B REPRESENTS STANDARD DEVIATION)

Dataset	**SGAI**	**SPSF-LAR**	**MRF**	**FCBF**	**mRMR**
spect	0.2997/0.0615	0.2994/0.0987	0.3110/0.0734	**0.2926/0.0724**	0.2994/0.0987
wdbc	**0.0281/0.0220**	0.3210/0.0289	0.0298/0.0299	0.0333/0.0314	0.3210/0.0289
spectf	**0.1235/0.0580**	0.2016/0.0322	0.2132/0.0227	0.2058/0.0228	0.2016/0.0322
promoter	0.2472/0.1448	0.2029/0.0901	**0.2009/0.1604**	0.2165/0.1691	0.2029/0.0901
infant	**0.0431/0.0070**	0.0446/0.0055	0.0457/0.0068	0.2596/0.0385	0.0446/0.0055
arcene	**0.1900/0.1101**	0.1980/0.1100	0.4600/0.067	0.2000/0.1563	0.1980/0.1100
dexter	**0.1067/0.0410**	0.2615/0.1761	0.5000/0	0.1500/0.0729	0.2615/0.1761
dorothea	0.0600/0.0275	0.0625/0.0283	0.090/0.0099	**0.0568/0.0223**	0.0625/0.0283
madelon	**0.1885/0.0225**	0.3088/0.0377	0.4005/0.0284	0.3341/0.0318	0.3088/0.0377
colon	0.1238/0.1554	**0.1043/0.1366**	0.1404/0.1689	0.1090/0.1396	**0.1043/ 0.1366**
prostate	0.0582/0.0502	0.0558/0.0617	**0.0482/0.0694**	0.0510/0.0515	0.0558/0.0617
leukemia	**0.0125/0.0395**	0.0265/0.0723	0.0143/0.0452	0.0150/0.0439	0.0265/0.0723
lung-cancer	0.0111/0.0234	**0/0**	0.0056/0.0176	0.0014/0.0087	**0/0**
breast-cancer	**0.0911/0.0448**	0.0945/0.0552	0.0975/0.0580	0.1008/0.0482	0.0945/0.0552
ovarian-cancer	0.0697/0.0515	0.0824/0.0556	0.0606/0.0452	**0.0577/0.0534**	0.0824/ 0.0556
hiva	0.0354/0.0017	**0.0351/0.0013**	0.0354/0.0017	**0.0351/0.0013**	**0.0351/0.0013**
average ranks	2.72	**2.56**	2.87	3.84	3.00
win/tie/loss	-	6/6/4	6/8/2	7/7/2	5/8/3