Industrial Engineering
& Management Systems
Vol 14, No 2, June 2015, pp.210-219
ISSN 1598-7248 | EISSN 2234-6473 |
http://dx.doi.org/10.7232/iems.2015.14.2.210

Classification of High Dimensionality Data through Feature Selection Using Markov Blanket

Junghye Lee, Chi-Hyuck Jun*
Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, Korea

(Received: May 4, 2015 / Revised: May 28, 2015 / Accepted: June 6, 2015)

ABSTRACT

A classification task requires an exponentially growing amount of computation time and number of observations as the variable dimensionality increases. Thus, reducing the dimensionality of the data is essential when the number of observations is limited. Often, dimensionality reduction or feature selection leads to better classification performance than using the whole number of features. In this paper, we study the possibility of utilizing the Markov blanket discovery algorithm as a new feature selection method. The Markov blanket of a target variable is the minimal variable set for explaining the target variable on the basis of conditional independence of all the variables to be connected in a Bayesian network. We apply several Markov blanket discovery algorithms to some high-dimensional categorical and continuous data sets, and compare their classification performance with other feature selection methods using well-known classifiers.

Keywords: Feature Selection, Classification, High Dimensionality Data, Markov Blanket

* Corresponding Author, E-mail: chjun@postech.ac.kr

1. INTRODUCTION

A classification problem is to predict a target variable of an observation on the basis of the features involved. When dealing with this problem, one of the most important things to consider is dimensionality reduction. If the number of features increases, the accuracy of classification generally increases also. However, this applies only when the number of observations is infinitely many. Exponential growth in the number of observations is required to accurately estimate a function for the target variable as the dimension increases. It is called the curse of dimensionality. In the actual data, however, since there are a finite number of observations, the accuracy of classification may decrease from the moment that the number of features exceeds a certain threshold because features, which are less relevant to the target variable, play a role in disturbing the classification in a finite number of observations. Successful reduction of dimensionality can achieve higher classification accuracy than using the entire features. Moreover, dimensionality reduction brings additional benefits such as reducing the time and memory complexity. Research on dimensionality reduction has been recognized as significant and it has been carried out very actively.

Two basic types of dimensionality reduction include feature extraction and feature selection. Feature extraction is transforming the existing features into a lower dimensional space. On the other hand, feature selection is selecting a subset of the existing features without a transformation. The representative techniques of feature extraction are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) which assume the linearity of the function. Also, nonlinear methods using kernels and other varieties are available such as ISO maps. Feature selection is divided into three categories – filter methods, wrapper methods, and embedded methods (Guyon and Elisseeff, 2003; Saeys *et al*., 2005).

Each method has its own advantages and disadvantages. In this paper, we focus on the filter methods which are relatively simple and fast in computation. Furthermore, these methods can be tested by any classifier since feature selection task is executed independently of the classifier.

The Markov blanket feature selection method belongs to a filter method. Koller and Sahami (1996) defined that the Markov blanket of a target variable is the minimal set of features conditioned on which all other features are independent of the target variable in a probabilistic graphical model. In other words, the Markov blanket of a target variable is the minimum information to explain the target variable fully. Based on this, the Markov blanket can be utilized as a feature selection method when the target variable is a class variable.

The usefulness of Markov blanket feature selection has been demonstrated in a few papers (Zeng *et al*., 2009), but the classification performance has not been reported extensively. In this paper, we compare three algorithms of Markov blanket discovery to test how they perform in classifying high-dimensional categorical and continuous data. These algorithms include Incremental Association Markov Blanket (IAMB) (Tsamardinos *et al*., 2003a), Max-Min Markov Blanket (MMMB) (Tsamardinos *et al*., 2003b), and HITON Markov Blanket (HITON-MB) (Aliferis *et al*., 2003a). Common classifiers are considered such as Naïve Bayes (NB), support vector machine (SVM), and *k*-nearest neighborhood (KNN). When comparing the classification performance, we also include two other feature selection methods: Correlation–based Feature Selection (CFS) (Hall, 1999) and two versions of Minimum Redundancy Maximum Relevance (MRMR) method (Ding and Peng, 2005).

In Section 2, Markov blanket is defined in the context of a probabilistic graphical model called a Bayesian network. Then in Section 3, we introduce three Markov blanket discovery algorithms for feature selection. Section 4 provides a description of two other feature selection methods which are compared with Markov blanket algorithms, and describes several classifiers which are used in this paper. In Section 5, we report the classification performance of each feature selection method combined with each classifier, which is applied to four categorical data sets and four continuous data sets. In Section 6 we conclude the paper with a summary of observations from the experiments.

2. BAYESIAN NETWORK AND MARKOV BLANKET

A Bayesian network is a probabilistic graphical model that compactly represents a joint probability distribution P over a set of random variables U via a directed acyclic graph (DAG) G. Its nodes represent random variables and the edges involve conditional dependencies between nodes. If the Markov condition property holds in a Bayesian network, then a node is independent from all nodes other than its descendants when conditioned on its parents (Pearl, 1988). Therefore, a Bayesian network consists of a qualitative part in the form of a DAG and a quantitative part in the form of conditional probabilities (Van Harmelen *et al*., 2008).

All Markov blanket discovery algorithms begin with two basic assumptions. The first is correctness of a conditional independence test, which means that we always obtain the correct result by the conditional independence test. The second assumption is faithfulness between a Bayesian network G and a joint distribution P, which indicates that every conditional independence entailed by the graph G and the Markov condition have to be presented in P (Fu and Desmarais, 2008; Fu and Desmarais, 2010; Pearl, 1988).

Now, the Markov blanket is defined formally as follows (Fu and Desmarais, 2010):

- **Definition 1 (Markov Blanket)**

Given the faithfulness assumption, from the perspective of the probability, the Markov Blanket of a target variable T, denoted by $MB(T)$, is the minimal set of variables conditioned on which all other variables F are independent of T. In the graphical perspective, the Markov blanket of T is the union of parent, child (PC), and parent of children, spouse (SP), nodes of T. For example, in Figure 1, the parent and child nodes of T are $PC(T) = \{A, B, C\}$, and the spouse node is $SP(T) = \{D\}$. So, the Markov blanket for T is $MB(T) = \{A, B, C, D\}$. It means that nodes E, F, and G are independent of T conditioned on $MB(T)$ (Fu and Desmarais, 2010).

3. MARKOV BLANKET DISCOVERY ALGORITHMS AS FEATURE SELECTION METHODS

In this section, three algorithms for Markov blanket discovery are introduced. In the algorithm, $(X \perp Y \mid \boldsymbol{Z})$ represents that X and Y are independent given a node set of $\boldsymbol{Z}$ and $dep(X, Y \mid \boldsymbol{Z})$ is the degree (or score) of the dependence between X and Y given $\boldsymbol{Z}$ which is p-value of a conditional independence test. In the case of cate-

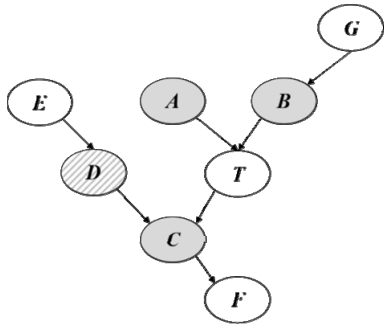

Figure 1. An example Bayesian network.

gorical variables, all of the Markov blanket discovery algorithms implement a G^2 conditional independence test (McDonald, 2009). On the other hand, in the case of continuous variables, they conduct a conditional independence test based on Fisher's z-transformation of the partial correlation coefficient.

3.1 Incremental Association Markov Blanket

The IAMB algorithm (Tsamardinos *et al*., 2003a) is the basic algorithm to discover the Markov blanket. Figure 2 is the pseudo code of IAMB. It is a grow-and-shrink approach that consists of two phases. In the first grow phase, nodes determined to be dependent on the target node are added to *MB* through the independence test (lines 2-6). In the next shrink phase, any node among *MB* determined to be independent of the target node is removed from *MB* (line 7-9).

Tsamardinos *et al*. (2003a) proved that IAMB satisfies soundness (correctness) under the faithfulness assumption. In order to achieve a reliable result from the algorithm, independence tests have to be correct, which means that they conclude (in)dependence if and only if the (in)dependence holds in *P*. However, IAMB has a drawback in terms of the data efficiency (Peña *et al*., 2007). IAMB is known to give a reliable result in discovering MB when the amount of instances is at least five times the degree of freedom in the test. In other words, IAMB requires that the number of instances increases exponentially according to the size of MB because the degree of freedom in the test is exponentially increasing in the size of the conditioning set, and the size of the conditioning set is the same as MB in IAMB.

3.2 Max-Min Markov Blanket

The MMMB algorithm (Tsamardinos *et al*., 2003b) tries to overcome the data inefficiency of IAMB while still being scalable given the faithfulness assumption. MMMB is also divided into two phases, but it takes the divide-and-conquer approach which is different from IAMB in terms of using topological information. At the first phase (called MMPC), parent and child nodes of *T* are identified, and then the spouse nodes of *T* are to be found in MMMB phase. Not all nodes, although determined to be dependent on the target node in the test, may be included into the MB. Figure 3 is the pseudo code of MMMB. However, as stated in Pena *et al*. (2007), the MMMB algorithm does not guarantee the correct output under the faithfulness, but it works well in practical applications. Compared to IAMB, MMMB is slow because MMPC considers every subset of the output as the conditioning set for the tests (line 4 in MMPC (T)).

3.3 HITON Markov Blanket

The HITON-MB algorithm (Aliferis *et al*., 2003a) is similar to MMMB in terms of data efficiency, sound-

```
IAMB(T)
/* add true positives to MB */
1   MB = ∅
2   repeat
3   Y = arg : max_{X∈(U\MB\{T})} dep(T, X|MB)
4   if T ⊥̸ Y|MB then
5       MB = MB ∪ {Y}
6   until MB does not change
/* remove false positives from MB */
7   for each X ∈ MB do
8       if T ⊥ X|(MB\{X}) then
9           MB = MB\{X}
10  return MB
```

Figure 2. IAMB algorithm.

```
MMPC(T)
/* add true positives to PC */
1   PC = ∅
2   repeat
3     for each X ∈ (U\MB\{T}) do
4       Sep[X] = arg : max_{X⊆PC} dep(T, X|Z)
5       Y = arg : max_{X∈(U\PC\{T})} dep(T, X|Sep[X])
6     if T ⊥̸ Y|Sep[Y] then
7       PC = PC ∪ {Y}
8   until PC does not change
/* remove false positives from PC */
9   for each X ∈ PC do
10    if T ⊥ X|Z for some Z ⊆ PC\{X} then
11      PC = PC ∪ {X}
12  return PC

MMMB(T)
/* add true positives to MB */
1   PC = MMPC(T)
2   MB = PC
3   CanMB = (PC ∪_{X∈PC} MMPC(X))\{T}
/* add more true positives to MB */
4   for each X ∈ CanMB\PC do
5     find any Z such that T ⊥ X|Z and T, X ∉ Z
6     for each Y ∈ PC do
7       if T ⊥̸ X|Z∪{Y} then
8         MB = MB ∪ {X}
9   return MB
```

Figure 3. MMMB algorithm.

```
HITON-PC(T)
1   PC = ∅
2   CanPC = U \ {T}
3   repeat
/* add the best candidate to PC */
4   Y = arg : max_{X∈CanPC} dep(T, X|∅)
5   PC = PC ∪ {Y}
6   CanPC => CanPC ∪ {Y}
/* remove false positives from PC */
7   for each X ∈ PC do
8     if T ⊥ X|Z for some Z ⊆ PC \ {X} then
9       PC = PC \ {X}
10  until CanPC is empty
11  return PC

HITON-MB(T)
/* add true positives to MB */
1   PC = HITON-MB(T)
2   MB = (PC ∪_{X∈PC} MMPC(X)) \ {T}
/* remove false positives from MB */
3   for each X ∈ MB do
4     for each Y ∈ PC do
5       if T ⊥ X|Z for some
          Z ⊆ {Y} ∪ (U \ {T, X, Y}) then
6         MB = MB \ {X}
7   return MB
```

Figure 4. HITON-MB algorithm.

ness, and time complexity. Like MMMB, HITON-MB takes a divide-and-conquer approach to identifying MB: first finding PC and then, finding SP. Figure 4 is the pseudo code of HITON-MB. The algorithm proceeds in the same manner as MMMB except that it combines addition and removal steps in a same loop for the purpose of removing false positives as early as possible to make the conditioning set small. However, Pena *et al*. (2007) proved that HITON-MB does not guarantee the correct output under the faithfulness. Since HITON-MB uses the topology of G in the same manner as MMMB, it returns similar results with MMMB.

4. OTHER FEATURE SELECTION METHODS AND CLASSIFIERS UNDER CONSIDERATION

In this section, first, we briefly describe the other feature selection methods which will be compared with Markov blanket feature selection methods described in Section 3, and then explain the commonly used classifiers which are adopted in this paper for the experiment.

4.1 Other Feature Selection Methods

The feature selection methods to be described in this section are selected because these are multivariate filter methods which are popular in the area of bioinformatics. Just like the Markov blanket feature selection methods, these feature selection methods can be applied to both categorical and continuous data.

4.1.1 Correlation-based Feature Selection

CFS considers every subset of all features, which is based on the following philosophy: a good feature subset contains features which are highly correlated with the target (class) variable and not redundant between them (Hall, 1999). Consider a subset S of all features, which consists of k features $f_1, \cdots, f_k$. Let $r_{f_i f_j}$ be the correlation coefficient of f_i and f_j and let r_{cf_i} be the correlation coefficient between the target (class) variable and f_i. Then, the CFS is to find the subset having the following maximum score.

$$CFS = max_s \left[\frac{r_{cf_1} + r_{cf_2} + \cdots + r_{cf_k}}{\sqrt{k + 2\left(r_{f_1 f_2} + \cdots + r_{f_i f_j} + \cdots + r_{f_{k-1} f_k}\right)}} \right] \quad (1)$$

CFS can be run on Weka (Hall *et al*., 2009) with a best first search strategy. Like the greedy hill climbing, the best first search strategy moves through the search space by making local changes to the current feature subset. However, unlike the hill climbing, if the path being explored begins to look less promising, the best first search can back-track to a more promising previous subset and continue the search from there.

4.1.2 Minimum Redundancy Maximum Relevance

Ding and Peng (2005) developed the MRMR method which ranks features considering their relevance to the class variable and redundancy within features simultaneously. Top ranked features have larger relevance to the class variable and smaller redundancy within features, and they are regarded as more significant than others. Due to the ranking process, MRMR provides the right of choice for the number of features. MRMR method has two kinds of schemes to search for the next feature depending on the data type.

For categorical data features, the relevance of a feature to the class variable c is evaluated by the mutual information value between a feature and the class variable, which is denoted by $I(f_i, c)$. Mutual Information Difference (MID) and Mutual Information Quotient (MIQ) are defined, respectively, by

$$MID = max_{f_i \in \Omega_s} \left[I(f_i, c) - \frac{1}{|S|} \sum_{f_j \in S} I(f_i, f_j) \right] \quad (2)$$

$$MIQ = max_{f_i \in \Omega_s} \left[I(f_i, c) \Big/ \frac{1}{|S|} \sum_{f_j \in S} I(f_i, f_j) \right] \quad (3)$$

where the second term in the bracket is the average of all mutual information values between feature f_i and other features in S which represents the redundancy of f_i.

For continuous data features, the F-statistic is used as a measure of relevance between a feature and the class variable, which has the following form (Ding and Peng, 2005; Ding, 2002).

$$F(f_i, c) = \left[\Sigma_k n_k (\bar{v}_{ik} - \bar{v}_i)^2 / (K-1) \right] / \sigma^2 \quad (4)$$

where $\bar{v}_i$ is the average across all observations in f_i, $\bar{v}_{ik}$ is the average of f_i within the k-th class *(k = 1, ⋯, K)*, and $\sigma^2 = \left[\Sigma_k (n_k - 1) \sigma_k^2 \right] / (n - K)$ is the pooled variance (n_k and σ_k^2 are the size and the variance of the k-th class, respectively). For $K = 2$, the F statistic will reduce to the t statistic, with the relation $F = t^2$. On the other hand, as a measure of redundancy, the absolute value of Pearson correlation coefficient of f_i and f_j, which is denoted by $c(f_i, f_j)$, is chosen. Hence the F-test correlation difference (FCD) and the F-test correlation quotient (FCQ) can be defined as follows.

$$FCD = max_{f_i \in \Omega_s} \left[F(f_i, c) - \frac{1}{|S|} \sum_{f_j \in S} c(f_i, f_j) \right] \quad (5)$$

$$FCQ = max_{f_i \in \Omega_s} \left[F(f_i, c) \Big/ \frac{1}{|S|} \sum_{f_j \in S} c(f_i, f_j) \right] \quad (6)$$

4.2 Classifiers under Consideration

Classifiers described in this section are selected since they are commonly used and easy to implement. The first two classifiers are parametric methods, and the last classifier is nonparametric. Because the model complexity of nonparametric methods is relatively high, the KNN may cause an over-fitting problem.

4.2.1 Naïve Bayes

The NB classifier is a simplified version of evaluating the posterior probability of each class for the classification purpose (Zhang, 2004). Suppose an observed instance consists of p-dimensional feature $f = (f_1, f_2, \cdots, f_p)$. Then, using the Bayes' rule, the posterior probability of j-th class, denoted by $p(c_i | f_1, f_2, \cdots, f_p)$, can be calculated:

$$p(c_j | f_1, f_2, \cdots, f_p) \propto p(f_1, f_2, \cdots, f_p | c_j) p(c_j) \quad (7)$$

where $p(f_1, f_2, \cdots, f_p | c_j)$ is the likelihood and $p(c_j)$ is the prior probability of each class. The goal of the Bayes' rule is to find the decision boundary that every instance is assigned to the class with the highest posterior probability. The key assumption of the naïve Bayes is that conditioned on the class, the distribution of input features $f_1, f_2, \cdots, f_p$ is independent. Due to the assumption, the likelihood can be expressed in a product form:

$$p(f_1, f_2, \cdots, f_p | c_j) \propto \prod_{k=1}^{p} p(f_k | c_j) \quad (8)$$

Although it is simple and straightforward to implement, the NB is often well-performed, more so than the sophisticated classification methods.

4.2.2 Support Vector Machine

Vapnik and Cortes (1995) first invented the SVM. Its performance has been increasingly recognized and it has become one of the most powerful classification methods. Under p-dimensional input feature space, for the two-class problem, SVM seeks the two parallel hyperplanes which maximize the distance (or margin) between them and the *(p-1)*-dimensional hyperplane placed in the middle of the two parallel hyperplanes plays the role of a discriminant function. SVM is based on the hypothesis that the larger the margin between these parallel hyperplanes, the better the performance of the classifier will be. These hyperplanes can be derived by solving optimally a quadratic programming. One advantage of SVM is to consider a nonlinearity of data by introducing a variety of kernel functions. In this study, however, we do not use any kernel function.

4.2.3 *k*-nearest Neighborhood

The k-nearest neighborhood (KNN) method was first introduced by Fix and Hodges (1989). Since it is a non-parametric method for classifying an instance based on k closest instances, a similarity measure (Euclidean distance or others) is calculated between all pairs of instances in a dataset. Whenever a new data point has to be classified, its k-closest neighbors are found from the training data (k being the number of neighbors) by sorting the distance matrix. The most dominant class label in the set of neighbors is finally assigned to the new data. The best choice of k depends upon the data. Larger values of k reduce the effect of noise on the classification but make boundaries between classes less distinct. Generally, k is often chosen close to the square root of the data size (Fukunaga, 1990). In this paper, we change k from 1 to 10, and then the best result of k is recorded.

5. EXPERIMENTS

This section describes our experiments on four categorical data sets and four continuous data sets, and reports their classification performance results, at first, using three Markov Blanket discovery methods introduced in Section 3 and three other feature selection methods in Section 4. Then, three classifiers, including naïve Bayes, support vector machine, and k-nearest

neighborhood, are applied to each selected feature set. Therefore, 18 classification models are executed for each data set basically. To avoid the bias between the training and test data, we compare each model by averaging 5 runs of 5-fold cross validation. All other algorithms except CFS are run in Matlab. CFS is run in Weka with a best first search strategy. For MB discovery algorithms, we used the Matlab version of Causal Explorer toolkit (Aliferis *et al*., 2003b). However, MMMB is not available for continuous data. We experimented on the MB feature selection methods with different significant levels. The significant level is used for implementing the conditional independence test, and it may result in a different output of the selected features. Since the MRMR method is based on ranking process, the number of features needs to be fixed beforehand. Two or three levels are considered here depending on the number of all features. For example, in the Audiology data set having a total of 69 features; 20, 10, and 5 features are selected in MRMR to keep the balance with the number of selected features in MB.

5.1 Categorical Data

For categorical data, three data sets, Audiology, Promoter, and Splice, were selected from the UCI repository of machine learning; and the other data set, Lung Cancer, is from Causality Workbench repository (Guyon *et al*., 2011). The information about the data sets is summarized in Table 1. The Audiology data set contains 69 features and one class variable divided into 24 classes, which is to predict the auditory state. The Promoter data set is to determine whether it is a promoter or not, using the information of gene sequences. The Splice data set is a type similar to the Promoter data set. The Lung Cancer data set is to predict lung cancer, using generic health status variables such as smoking and fatigue.

Tables 2-5 show the classification result of each data set, which includes the number of features selected, classification accuracies of test data and those of training data (numbers in parenthesis) in percentage. The best performing feature selection method for each classifier is marked in bold numbers. In Table 2, IAMB with 5% significance level outperforms other feature selection methods for all classifiers with a significantly large gap. This result is remarkable when considering that this data set contains a large number of classes as many as 24. Using only 8 features, IAMB records the best accuracy among the feature selection methods, and it even has better performance than the entire features. For this data set, MMMB and HITON-MB do not provide any MB, so the results are not reported here. This tells us about some drawbacks of MMMB and HITON-MB although they use topology information differently from IAMB.

Tables 3, 4 and 5 show the similar results for Promoter, Splice and Lung Cancer data sets, respectively. In Tables 3-5, the results based on the MB feature selection methods are generally well-performed. Sometimes

Table 1. Categorical data sets for experiments

Data set	# features	# observations	# classes
Audiology	69	226	24
Promoter	57	106	2
Splice	60	200	2
Lung Cancer	143	100	2

Table 2. Performance comparison for Audiology data set

Feature selection		# features	Accuracies in percentage		
			NB	SVM	KNN
All features		69	63.60(77.24)	78.40(99.12)	59.91(99.54)
IAMB	0.01	13	83.74(87.93)	95.76(100.00)	89.20(100.00)
	0.05	8	**84.25(88.69)**	**96.18(100.00)**	**94.08(100.00)**
CFS		15	64.30(71.13)	64.04(72.98)	63.34(74.81)
MRMR-MID		5	68.94(76.21)	69.02(78.35)	68.69(73.33)
		10	68.08(79.36)	64.36(86.41)	58.42(90.30)
		20	71.82(85.26)	74.98(97.51)	65.88(98.66)
MRMR-MIQ		5	61.06(66.10)	55.97(62.55)	57.57(61.39)
		10	63.57(69.83)	61.20(70.95)	58.51(67.15)
		20	71.88(81.29)	75.68(93.30)	65.06(96.49)

Table 3. Performance comparison for promoter data set

Feature selection		# features	Accuracies in percentage		
			NB	SVM	KNN
All features		57	92.15 (98.95)	72.33 (100.00)	77.91 (82.39)
IAMB	0.01	4	92.85 (94.30)	69.85 (71.63)	85.69 (97.65)
	0.05	3	91.25 (94.44)	71.45 (71.32)	**86.02 (96.76)**
MMMB	0.01	24	90.73 (97.67)	**76.65 (96.54)**	73.64 (82.70)
	0.05	53	87.51 (98.62)	75.05 (100.00)	80.73 (80.92)
HITON-MB	0.01	24	91.31 (97.90)	73.91 (96.46)	74.56 (85.22)
	0.05	53	89.85 (98.72)	73.41 (100.00)	76.05 (88.91)
CFS		6	95.27 (95.47)	68.27 (73.25)	83.58 (100.00)
MRMR-MID		5	**95.29 (96.21)**	67.16 (72.77)	85.91 (96.13)
		20	95.04 (96.85)	72.85 (93.44)	83.52 (76.45)
		50	92.47 (98.78)	70.85 (100.00)	73.56 (80.88)
MRMR-MIQ		5	93.13 (96.52)	68.96 (72.79)	85.02 (96.83)
		20	94.33 (97.44)	71.07 (93.90)	76.20 (82.34)
		50	90.75 (98.74)	74.82 (100.00)	73.25 (86.62)

Table 4. Performance comparison for splice data set

Feature selection		# features	Accuracies in percentage		
			NB	SVM	KNN
All features		60	92.67 (99.89)	80.00 (100.00)	97.77 (98.68)
IAMB	0.01/0.05	4	93.86 (97.01)	81.19 (84.60)	97.90 (98.86)
MMMB/ HITON-MB	0.01	14	96.44 (99.10)	86.74 (97.62)	98.23 (98.99)
	0.05	24	95.05 (99.77)	84.98 (100.00)	98.34 (99.10)
CFS		8	95.45 (99.09)	85.12 (94.16)	98.43 (98.90)
MRMR-MID		5	96.04 (99.20)	80.05 (87.63)	**98.56 (98.90)**
		10	96.24 (99.21)	**87.71 (96.83)**	97.54 (98.53)
		20	**97.23 (99.38)**	81.62 (100.00)	97.89 (98.73)
MRMR-MIQ		5	93.88 (98.07)	82.97 (89.60)	98.43 (98.85)
		10	95.47 (99.34)	87.54 (96.34)	98.24 (98.94)
		20	97.23 (99.72)	84.00 (100.00)	98.35 (99.13)

MB does not return the best accuracy but there is not a big difference. We note that MMMB and HITON-MB discover almost the same MB. Even though the MB feature selection methods do not guarantee to provide the best solution in all data sets, their usefulness is still evident since the selection method reduces the number of features significantly while maintaining the good performance. Among these MB discovery algorithms, IAMB generally performs well for all these categorical data sets.

5.2 Continuous data

For continuous data, all data sets are from Kent Ridge Bio-medical repository (Li and Liu, 2002). The information about the data sets is in Table 6. All continuous data sets are microarray gene expression data, which are high-dimensional in features having relatively small number of observations. The AML/ALL data set is to predict the presence of acute myeloid leukemia or acute lymphoblastic leukemia. Colon Cancer, Prostate Cancer and Ovarian Cancer data sets are to predict colon cancer, prostate cancer and ovarian cancer of patients, respectively.

Tables 7-10 show the performance result for each data set of each classification model, which includes the number of features selected, classification accuracy of the test data and accuracy of the training data (in parenthesis). The best performing feature selection method for each classifier is marked in bold numbers.

Table 5. Performance comparison for Lung cancer data set

Feature selection		# features	Accuracies in percentage		
			NB	SVM	KNN
All features		143	93.23 (99.00)	87.81 (100.00)	97.40 (98.49)
IAMB	0.01/0.05	6	93.81 (98.89)	**94.24 (94.85)**	**98.50 (99.72)**
MMMB	0.01	5	92.60 (98.88)	93.41 (95.95)	95.82 (96.16)
	0.05	5	91.60 (98.35)	91.03 (95.55)	93.54 (94.06)
HITON-MB	0.01	4	90.81 (97.80)	90.58 (92.51)	89.87 (90.18)
	0.05	8	94.01 (98.80)	88.21 (96.65)	95.48 (95.99)
CFS		17	**96.00 (99.77)**	89.62 (99.64)	97.83 (98.72)
MRMR-MID		5	93.01 (97.82)	90.78 (94.32)	93.80 (94.27)
		10	95.03 (99.00)	89.19 (96.81)	96.87 (97.41)
MRMR-MIQ		5	91.00 (96.99)	90.81 (92.35)	91.39 (91.60)
		10	91.81 (98.74)	89.03 (93.54)	94.97 (96.25)

Table 6. Continuous data sets for experiments

Data set	# features	# observations	# classes
AML/ALL	7129	72	2
Colon Cancer	2000	62	2
Prostate Cancer	12600	102	2
Ovarian Cancer	15114	253	2

Table 7. Performance comparison for AML/ALL data set

Feature selection		# features	Accuracies in percentage		
			NB	SVM	KNN
All features		7129	**98.89 (100.00)**	96.36 (100.00)	98.63 (99.26)
IAMB	0.01	62	88.89 (91.33)	98.32 (100.00)	97.38 (98.52)
	0.05	63	89.17 (92.17)	98.06 (100.00)	98.18 (99.04)
HITON-MB	0.01	4	98.61 (100.00)	98.04 (100.00)	99.88 (99.87)
	0.05	6	97.22 (98.97)	96.93 (100.00)	99.94 (100.00)
CFS		44	98.33 (99.78)	**100.00 (100.00)**	**100.00 (100.00)**
MRMR-FCD		5	94.44 (96.24)	90.17 (95.69)	99.02 (99.43)
		60	95.56 (96.87)	97.14 (100.00)	99.82 (99.94)
MRMR-FCQ		5	90.28 (92.34)	97.18 (100.00)	98.48 (99.12)
		60	96.11 (97.38)	97.83 (100.00)	99.41 (99.70)

It is observed in Tables 7-10 that the MB feature selection methods perform quite well for these continuous data sets when combined with suitable classifiers. It seems that IAMB extracts more features than HITON-MB for these continuous data sets. For the AML/ALL data set, the CFS method shows the best performance when combined with the classifier SVM or KNN, but the MB feature selection methods show relatively good results. For this data set, HITON-MB having 1% significance level produce better performance than any other feature selection methods when combined with the classifier NB. It should be noted that IAMB achieves the best performance for the Colon Cancer data set using only three features out of two thousands. For the Prostate and Ovarian Cancer data sets, the MB feature selection methods generally produce better performance. When comparing the CFS method with the MRMR method, the former generally performs better than the latter in these data sets.

6. CONCLUSION

We have shown that Markov blanket discovery algorithms can be utilized as feature selection methods by constructing a minimal set of features from a Bayesian network formed by the whole variables. Moreover, Mar-

Table 8. Performance comparison for Colon Cancer data set

Feature selection		# features	Accuracies in percentage		
			NB	SVM	KNN
All features		2000	57.42 (90.75)	83.82 (100.00)	95.04 (97.15)
IAMB	0.01	3	**88.71 (98.34)**	**86.82 (92.75)**	**97.97 (98.93)**
	0.05	55	62.58 (81.35)	86.49 (100.00)	94.52 (96.82)
HITON-MB	0.01	2	84.19 (97.78)	85.54 (89.12)	94.53 (96.84)
	0.05	7	87.10 (98.99)	86.56 (93.96)	95.65 (97.55)
CFS		6	52.58 (78.88)	69.77 (76.39)	93.80 (96.58)
MRMR- FCD		5	84.52 (97.97)	83.44 (88.38)	96.80 (98.21)
		50	83.55 (97.30)	81.97 (100.00)	95.05 (97.11)
MRMR- FCQ		5	88.39 (98.01)	83.62 (90.17)	97.03 (98.29)
		50	87.42 (99.43)	83.03 (100.00)	96.70 (98.09)

Table 9. Performance comparison for Prostate Cancer data set

Feature selection		# features	Accuracies in percentage		
			NB	SVM	KNN
All features		12600	62.94 (81.11)	91.18 (100.00)	97.11 (98.34)
IAMB	0.01	88	83.73 (90.44)	96.65 (100.00)	98.21 (99.02)
	0.05	89	83.53 (90.14)	**98.23 (100.00)**	96.88 (98.16)
HITON-MB	0.01	5	92.55 (95.71)	93.50 (97.21)	99.02 (99.44)
	0.05	7	**94.71 (98.78)**	93.54 (96.62)	98.47 (99.09)
CFS		27	93.14 (97.27)	97.42 (100.00)	**99.87 (99.96)**
MRMR – FCD		5	91.96 (94.98)	93.16 (93.38)	98.39 (99.15)
		30	93.73 (96.74)	91.58 (100.00)	98.28 (99.00)
		90	92.35 (95.47)	93.15 (100.00)	98.85 (99.32)
MRMR – FCQ		5	88.63 (92.11)	93.51 (96.62)	98.34 (99.04)
		30	93.73 (95.99)	94.90 (100.00)	99.06 (99.43)
		90	94.12 (98.77)	95.90 (100.00)	98.41 (99.11)

Table 10. Performance comparison for Ovarian Cancer data set

Feature selection		# features	Accuracies in percentage		
			NB	SVM	KNN
All features		15046	91.07 (95.43)	99.84 (100.00)	98.97 (99.40)
IAMB	0.01	30	**99.64 (100.00)**	**100.00 (100.00)**	**100.00 (100.00)**
	0.05	224	93.36 (97.52)	**100.00 (100.00)**	99.78 (99.87)
HITON-MB	0.01	7	99.53 (100.00)	**100.00 (100.00)**	**100.00 (100.00)**
	0.05	9	98.97 (99.33)	99.92 (100.00)	**100.00 (100.00)**
CFS		28	99.60 (100.00)	**100.00 (100.00)**	**100.00 (100.00)**
MRMR- FCD		5	96.99 (98.23)	97.00 (97.02)	98.87 (99.34)
		10	96.13 (97.89)	97.23 (97.61)	99.05 (99.45)
		30	97.94 (99.02)	98.97 (99.84)	99.71 (99.84)
MRMR- FCQ		5	96.92 (98.04)	97.94 (98.26)	99.31 (99.61)
		10	98.02 (98.98)	98.81 (99.66)	99.93 (99.98)
		30	98.97 (99.35)	99.92 (100.00)	99.97 (100.00)

kov blanket discovery algorithms are shown to be competitive in the classification performance as compared to other popular feature selection methods in the experiments with categorical and continuous data sets. Among these MB discovery algorithms, the IAMB algorithm, which is simplest, generally performs quite well for all categorical and continuous data sets considered.

ACKNOWLEDGEMENTS

This research was supported by a grant of the Korea Health technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (Grant Number: HI13C-0790-010013).

REFERENCES

Aliferis, C. F., Tsamardinos, I., and Statnikov, A. (2003a), Hiton: a novel Markov blanket algorithm for optimal variable selection, *American Medical Informatics Association Annual Symposium Proceedings*, 21-25.

Aliferis, C. F., Tsamardinos, I., Statnikov, A., and Brown, L. E. (2003b), Causal explorer: a causal probabilistic network learning toolkit for biomedical discovery, *METMBS Conference*, **3**, 371-376.

Ding, C. (2002), Analysis of gene expression profiles: class discovery and leaf ordering, *Proceedings of the 6th Annual International Conference on Computational Biology*, 127-136.

Ding, C. and Peng, H. (2005), Minimum redundancy feature selection from microarray gene expression data, *Journal of Bioinformatics and Computational Biology*, **3**(2), 185-205.

Fix, E. and Hodges, J. L. (1989), Discriminatory analysis-nonparametric discrimination: consistency properties, *International Statistical Review*, **57**(3), 238-247.

Fu, S. and Desmarais, M. C. (2008), Tradeoff analysis of different Markov blanket local learning approaches, in: Washio, T. Suzuki, E., Ting, K. M., Inokuchi, A. (Eds.), *Advances in Knowledge Discovery and Data Mining*, Springer, Osaka, 562-571.

Fu, S. and Desmarais, M. C. (2010), Markov blanket based feature selection: a review of past decade, *Proceedings of the World Congress on Engineering*, **1**, 321-328.

Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition*, second ed. Academic Press, San Diego.

Guyon, I. and Elisseeff, A. (2003), An introduction to variable and feature selection, *Journal of Machine Learning Research*, **3**, 1157-1182.

Guyon, I., Aliferis, C., Cooper, G., Eisseeff, A., Pellet, J. P., Spirtes, P., and Statnikov, A. (2011), *Causality workbench*, in: Illari, P.M., Russo, F., Williamson, J. (Eds.), Causality in the Sciences. Oxford University Press, Oxford.

Hall, M. A. (1999), *Correlation-based feature selection for machine learning*, Unpublished doctoral dissertation, University of Waikato, Hamilton, New Zealand.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009), The WEKA data mining software: an update, *ACM SIGKDD Explorations Newsletter*, **11**(1), 10-18.

Koller, D. and Sahami, M. (1996), Toward optimal feature selection, *Proceedings of 13th International Conference on Machine Learning*, **45**(2), 211-232.

Li, J. and Liu, H. (2002), Kent ridge bio-medical data set repository, *Institute for Infocomm Research*, http://sdmc.lit.org.sg/GEDatasets/Datasets.html.

McDonald, J. H. (2009), *Handbook of Biological Statistics*, second ed. Sparky House Publishing, Baltimore.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, second ed., Morgan Kaufmann Publishers, Inc., San Francisco.

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

Saeys, Y., Inza, I., and Larrañaga, P. (2005), A review of feature selection techniques in Bioinformatics, *Bioinformatics*, **23**(19), 2507-2517.

Tsamardinos, I., Aliferis, C. F., and Statnikov, A. (2003), Algorithms for large scale Markov blanket discovery, *American Association for Artificial Intelligence*, 376-381.

Tsamardinos, I., Aliferis, C. F., and Statnikov, A. (2003b), Time and sample efficient discovery of Markov blankets and direct causal relations, *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 673-678.

Van Harmelen, F., Lifschitz, V., and Porter, B. (2008), *Handbook of Knowledge Representation*, first ed. Elsevier, Amsterdam.

Vapnik, V. and Cortes, C. (1995), Support-vector networks, *Machine Learning*, **20**(3), 273-297.

Zeng, Y., Luo, J., and Lin, S. (2009), Classification using Markov blanket for feature selection, *IEEE International Conference on Granular Computing*, 743-747.

Zhang, H. (2004), The optimality of naive Bayes, *Proceedings of the 17th International FLAIRS Conference*, **1**, 3-9.