Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Expert Systems with Applications

An International Journal

Editor-in-Chief
Binshan Lin

# An efficient multivariate feature ranking method for gene selection in high-dimensional microarray data

Junghye Lee[a,*], In Young Choi[b], Chi-Hyuck Jun[c]

[a] Department of Industrial Engineering, Ulsan National Institute of Science and Technology (UNIST), 50 UNIST-gil, Ulsan, Republic of Korea
[b] Department of Medical Informatics, College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seoul, Republic of Korea
[c] Department of Industrial and Management Engineering, Pohang University of Science and Technology (POSTECH), 77 Cheongam-ro, Pohang, Republic of Korea

ARTICLE INFO

*Keywords:*
Multivariate feature selection
Gene selection
High-dimensional data
Microarray data
Classification
Multiclass
Mixed-type data
Markov blanket
Ranking

ABSTRACT

Classification of microarray data plays a significant role in the diagnosis and prediction of cancer. However, its high-dimensionality (>tens of thousands) compared to the number of observations (<tens of hundreds) may lead to poor classification accuracy. In addition, only a fraction of genes is really important for the classification of a certain cancer, and thus feature selection is very essential in this field. Due to the time and memory burden for processing the high-dimensional data, univariate feature ranking methods are widely-used in gene selection. However, most of them are not that accurate because they only consider the relevance of features to the target without considering the redundancy among features. In this study, we propose a novel multivariate feature ranking method to improve the quality of gene selection and ultimately to improve the accuracy of microarray data classification. The method can be efficiently applied to high-dimensional microarray data. We embedded the formal definition of relevance into a Markov blanket (MB) to create a new feature ranking method. Using a few microarray datasets, we demonstrated the practicability of MB-based feature ranking having high accuracy and good efficiency. The method outperformed commonly-used univariate ranking methods and also yielded the better result even compared with the other multivariate feature ranking method due to the advantage of data efficiency.

## 1. Introduction

Recently many researchers have proposed that DNA microarray technology is able to measure the expression levels of thousands of genes simultaneously and contribute for diagnosis and development of the proper treatment plan for patients. Microarray, which is one specific technology to obtain gene expression data, allows monitoring of thousands of genes in parallel and produce enormous valuable data.

Gene expression microarray data can diagnose cancer and helps to find out the appropriate treatment plan based on the characteristics of patient. Traditionally cancer is diagnosed based on its morphological and clinical features. Gene information is well known as it can define the phenotype or the symptom of cancer disease and can improve disease progression and outcome prediction (Alon et al., 1999; Golub et al., 1999; Ross et al., 2000). Cancer classification using microarray data helps doctors to suggest a care plan in an efficient way, which can improve the quality of lives (Mabu, Prasad, & Yadav, 2020).

Popular classification methods such as logistic regression (Cox, 1958), Fisher's discriminant analysis (Fisher, 1936), *k*-nearest neighborhood (*k*-NN) (Fix & Hodges, 1989), and support vector machine (SVM) (Cortes & Vapnik, 1995) can be applied to this problem, but one of the major problems not doing so in microarray data is its high dimensionality (>tens of thousands) compared to a small number of observations (<tens of hundreds). Given a sample size, classifiers become complicated and degraded when the dimensionality of the input feature space is very large (Trunk, 1979).

Microarray data consist of a number of features, but in most cases, only a fraction of genes is important (Ben-Dor et al., 2000). Therefore, feature selection (i.e., gene selection) is necessarily required to identify the important genes which help classify samples effectively, and many gene selection-related studies have been introduced in the past (Kononenko, 1994; Ben-Dor et al., 2000; Guyon, Weston, Barnhill, & Vapnik, 2002; Ding & Peng, 2005; Chandra & Gupta, 2011; Wang, Zhou, Yi, & Kong, 2014; Abdulqader, Abdulazeez, & Zeebaree, 2020). However, due to the burden of time and memory complexity, univariate feature ranking methods are commonly used in feature selection of gene

[*] Corresponding author.
*E-mail addresses:* junghyelee@unist.ac.kr (J. Lee), iychoi@catholic.ac.kr (I.Y. Choi), chjun@postech.ac.kr (C.-H. Jun).

https://doi.org/10.1016/j.eswa.2020.113971
Received 24 March 2020; Received in revised form 25 August 2020; Accepted 2 September 2020
Available online 20 September 2020

expression data (Lê Cao, Bonnet, & Gadat, 2009); existing methods to find an optimal subset of features and multivariate feature ranking methods considering redundancy among features require computationally expensive search strategy, even using heuristic search methods and pairwise comparison between two features respectively. Univariate feature ranking methods have advantages of being fast and simple but those are vulnerable in accuracy because all redundant features are high-ranked or low-ranked together. As important as selecting the important features, it is crucial to construct a mutually exclusive important feature set in cancer classification because it leads directly to cost savings; researchers would like to see information about cancer with only a few genes due to cost issues.

The main purpose of this study is to provide a new multivariate feature ranking method having low complexity and high accuracy. The new method called Markov Blanket (MB) Ranking is basically based on an MB but has been modified to improve the classification accuracy of gene expression data. We embed the formal definition of relevance into the MB to make it a multivariate ranking method, which simultaneously considers relevance to the target and redundancy among features within reasonable time and memory complexity.

This paper is organized as follows. Related work is briefly introduced in Section 2. Section 3 presents the theoretical background to explain the proposed method in Section 4. Section 5 analyzes comparison results of the proposed method and benchmark methods on the several microarray datasets for classification in terms of prediction accuracy and computational efficiency. Finally, we conclude with the summary of this study including contributions and limitations in Section 6.

## 2. Related work

A number of gene selection methods have been introduced to select important genes for disease prediction and diagnosis. Gene selection approaches can be classified into five categories depending on the combination with the prediction model (Guyon & Elisseeff, 2003; Saeys, Inza, & Larrañaga, 2007; Ang, Mirzal, Haron, & Hamed, 2015; Manikandan & Abirami, 2018; Vanjimalar, Ramyachitra, & Manikandan, 2018): filter (Duch, Wieczorek, Biesiada, & Blachnik, 2004; Lazar et al., 2012), wrapper (Ruiz, Riquelme, & Aguilar-Ruiz, 2006; Wang et al., 2017), embedded (Hoque, Ahmed, Bhattacharyya, & Kalita, 2016), hybrid (Hsu, Hsieh, & Lu, 2011; Raweh, Nassef, & Badr, 2018; Almugren & Alshamlan, 2019), and ensemble (Shen, Diao, & Su, 2012). Filter methods employ the measure of feature relationship such as correlation (Van't Veer et al., 2002), consistency, distance (Kononenko, 1994), relevance (Wang et al., 2017), dependency, redundancy, and mutual information (MI) (Battiti, 1994; Brown, 2009; Peng, Long, & Ding, 2005; Devi Arockia Vanitha, Devaraj, & Venkatesuluc, 2015). The methods use a specific measure to identify the important features (Fulcher, 2008). They can be combined with any prediction models because their feature selection step is independent of model fitting. Wrapper methods such as stepwise regression iteratively perform feature selection in the direction of improving the quality criteria of a particular model, such as accuracy and error rate (Kohavi & John, 1997), and may include heuristic search methods, such as genetic algorithms and particle swarm optimization, to find an optimal subset of features. The wrappers are generally more accurate than the filters but are more complex and slower. The embedded methods identify the features that have great impact on the accuracy of the model during model fitting. Hybrid methods and ensemble methods are variants of the abovementioned three types of methods; the former ones are the combination of filter and wrapper methods and the latter ones aggregate the results of several feature subsets.

Wrapper, embedded, hybrid, and ensemble methods are generally more accurate than filter methods when an appropriate prediction model is given. However, they are more complex and slower than filter methods because prediction model learning is required for every candidate feature subset (Lamba, Munjal, & Gigras, 2018). Filter methods focused on this study are relatively simple and fast in computation. Furthermore, these methods can be tested by any prediction models since the feature selection task is executed independently to the prediction task. Therefore, features selected by the filter methods might be more general than those of the other methods.

Filters can be also categorized according to the type of output: feature subset selection and feature ranking (Bolón-Canedo, Sánchez-Maroño, & Alonso-Betanzos, 2015). Feature subset selection methods generate a subset of features that collectively have good predictive capability. They are based on the assumption that a given feature may have better predictive power when combined with some other attributes, compared to when used by itself. Feature subset selection provides one optimal subset of features, and thus users need not to consider the number of features to be selected (i.e., cutoff). On the other hand, feature ranking methods rank features by scoring each feature according to a particular method, then selecting features based on their scores. Feature ranking provides the rank of features and this requires the user's decision about the cutoff, but this enables feature ranking to have many choices to construct a prediction model depending on the cutoff instead. Usually, feature subset selection has a higher time complexity than feature ranking because feature subset selection can suffer from an inevitable problem caused by searching through feature subsets required in the subset generation step; the search space is $O(2^d)$. Although there exist various heuristic search strategies such as greedy sequential search, best-first search, and genetic algorithm (Liu, Motoda, & Dash, 1998), most of them still incur time complexity $O(d^2)$, which prevents them from scaling well to datasets containing tens of thousands of features.

Among them, most commonly used methods in gene selection are filters and especially feature ranking methods (García, Sánchez, Cleofas-Sánchez, Ochoa-Domínguez, & López-Orozco, 2017) due to their low complexity, and these feature ranking methods can be grouped again into two approaches: univariate (without considering redundancy) and multivariate (with considering redundancy) (Tang, Alelyani, & Liu, 2014; Liao et al., 2015). The former approach calculates the score between each feature and the target using $t$ statistic (Liu, Li, & Wong, 2002) and $\chi^2$ statistic (Liu & Setiono, 1995; Liu et al., 2002; García et al., 2017), whereas the latter approach captures the relationship among features though employing the abovementioned measures (Sun et al., 2019) or clustering algorithms (Chen, Zhang, & Gutman, 2016)

Most widely used and well-known feature ranking methods are described in detail as follows. Note that some of the studies will be used as benchmark methods for comparative experiments.

1) Effective Range Based Gene Selection (ERGS) (Chandra & Gupta, 2011): Let $X = \{X_i\}$ be the feature set $R^d$, $i = 1,2,\ldots,d$. $C = \{C_j\}$ $(j = 1,2,\ldots,l)$ is the class labels. The class probability of $j$th class $C_j$ is $p_j$. For each class $C_j$ of the $i$th feature $X_i$, $\mu_{ij}$ and $\sigma_{ij}$ denote the mean and standard deviation of the $i$th feature $X_i$ for class $C_j$, respectively. Effective range $(R_{ij})$ of $j$th class $C_j$ for $i$th feature $X_i$ is defined by

$$R_{ij} = \left[r_{ij}^{-}, r_{ij}^{+}\right] = \left[\mu_{ij} - \left(1 - p_j\right)\gamma\sigma_{ij}, \mu_{ij} + \left(1 - p_j\right)\gamma\sigma_{ij}\right], \quad (1)$$

where $r_{ij}^{-}$ and $r_{ij}^{+}$ are the lower and upper bounds of the effective range, respectively. The prior probability of $j$th class is $p_j$. Here, the factor $(1 - p_j)$ is taken to scale down effect of class with high probabilities and consequently large variance. The value of $\gamma$ is determined statistically by Chebyshev inequality defined as

$$P\left(\left|X - \mu_{ij}\right| \geq \gamma\sigma_{ij}\right) \leq \frac{1}{\gamma^2} \quad (2)$$

which is true for all distributions. The value of $\gamma$ is set to 1.732 for the effective range which contains at least 2/3 of the data (Chandra & Gupta, 2011).

Overlapping area $(\mathrm{OA}_i)$ among classes of feature $X_i$ is computed by

$$\mathrm{OA}_i = \Sigma_{j=1}^{l-1}\Sigma_{k=j+1}^{l}\varphi_i(j,k), \tag{3}$$

where $\varphi_i(j,k)$ can be defined as

$$\varphi_i(j,k) = \begin{cases} r_{ij}^{+} - r_{ik}^{-} \;\; if \;\; r_{ij}^{+} > r_{ik}^{-} \\ 0 \quad otherwise. \end{cases} \tag{4}$$

In ERGS, for a given feature, the effective range of every class is first calculated. Then, the overlapping area of the effective ranges is calculated according to (3), and the area coefficient ($\mathrm{AC}_i$) is computed for each feature.

$$\mathrm{AC}_i = \frac{\mathrm{OA}_i}{\max\limits_j r_{ij}^{+} - \min\limits_j r_{ij}^{-}} \tag{5}$$

Next, the normalized area coefficient is regarded as the weight for every feature

$$w_i = 1 - \mathrm{NAC}_i \tag{6}$$

where $\mathrm{NAC}_i = \mathrm{AC}_i/\max(\mathrm{AC}_s)$ for $s = 1, \cdots, d$.

2) Improved Feature Selection Based on Effective Range (IFSER) (Wang et al., 2014): IFSER not only considers the overlapping areas of the features in different classes but also takes the including areas and the samples' proportion in overlapping and including areas into account.

$$\mathrm{IA}_i = \Sigma_{j=1}^{l-1}\Sigma_{k=j+1}^{l}\psi_i(j,k) \tag{7}$$

where $\psi_i(j,k)$ can be defined as

$$\psi_i(j,k) = \begin{cases} r_{ik}^{+} - r_{ik}^{-} \;\; if \;\; r_{ij}^{+} > r_{ik}^{-} \\ 0 \; otherwise. \end{cases} \tag{8}$$

The area coefficient ($\mathrm{AC}_i$) is computed for each feature

$$\mathrm{AC}_i = \frac{\mathrm{OA}_i + \mathrm{IA}_i}{\max\limits_j r_{ij}^{+} - \min\limits_j r_{ij}^{-}}. \tag{9}$$

Let $H_{ij}$, $G_{ij}$ denote samples' numbers of the $j$-th class in $\mathrm{OA}_i$, $\mathrm{IA}_i$ for each feature $X_i$, and $H_i = H_{ij}/n_j$, $G_i = G_{ij}/n_j$ are proportions of samples, where $n_j$ is the number of samples in the $j$th class. For all classes of each feature $X_i$, the normalized $H_i$ and $G_i$ for $s = 1, 2, \cdots, d$ can be obtained by

$$NH_i = 1 - \frac{H_i}{\max(H_s)}, \tag{10}$$

$$GH_i = 1 - \frac{G_i}{\max(G_s)}. \tag{11}$$

The weight of each feature $X_i$ is computed as

$$w_i = (1 - NAC_i)\bullet(NH_i + GH_i) \tag{12}$$

where $\mathrm{NAC}_i = \mathrm{AC}_i/\max(\mathrm{AC}_s)$ for $s = 1, \cdots, d$.

3) Pearson Coefficient Correlation (PCC): PCC is measure of the linear correlation between two variables $X$ and $Y$, giving a value between $+1$ and $-1$ inclusive, where 1 is total positive correlation, 0 is no correlation, and $-1$ is total negative correlation. It is widely used in the sciences as a measure of the degree of linear dependence between two variables and the formula of PCC $r$ is

$$r = r_{xy} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{n\Sigma x_i^2 - (\Sigma x_i)^2}\sqrt{n\Sigma y_i^2 - (\Sigma y_i)^2}}, \tag{13}$$

given n paired samples of $X$ and $Y$, $\{X_i, Y_i\}_{\{i=1\}}^{\{n\}}$.

4) Relief-F (Kononenko, 1994): Relief-F assigns a "relevance" weight to each feature. Randomly, a sample instance ($R$) is selected from data and the relevance values are updated based on the difference between the selected instance ($R$) and the nearest instances of the same ($H$) (called nearest hit) and different class ($M(C_j)$) (called nearest miss of class $C$). It gives more weight to features that discriminate the instance from neighbors of different classes. The weights are updated by considering average contribution of nearest misses $M(C_j)$. The average contribution also takes into account of prior probability of each class. The weight of $i$th feature $X_i$ is updated as follows.

$$w_i = w_i - \frac{\psi(X_i, R, H)}{m} + \Sigma_{C_j \neq C_R}\frac{p_j \bullet \psi(X_i, R, M(C_j))}{m} \tag{14}$$

The function $\psi(X_i, R, H)$ and $\psi(X_i, R, M(C))$ calculate the distance between sample instance ($R$) and nearest hit ($H$) and nearest misses $M(C_j)$ respectively; $m$ is the number of operations to update. There are notable works using Relief-F (García & Sánchez, 2015; Ke, Wu, Wu, & Xiong, 2018).

5) Information Gain (IG): IG is popularly used as attribute selection criteria in Decision Tree by (Quinlan, 2014). Liu et al. (2002) have used it as a gene selection criterion. For each feature $X_i$, IG is measured as

$$IG(X_i) = H(C) - H(C|X_i) \tag{15}$$

where

$$H(C) = -\Sigma p(C_j)\log p(C_j) \tag{16}$$

and

$$H(C|X_i) = -\Sigma_{x\in X_i} p(x)\Sigma p(C_j|x)\log p(C_j|x) \tag{17}$$

IG can be used only on discrete features and hence for numeric features discretization is necessary prior to computing IG. Features are selected based on the larger values of IG.

6) Gain Ratio (GR): To overcome the problem of IG that IG measure is biased towards features with a large number of values, GR is developed as an attempt to correct for this by normalizing IG by intrinsic information (IntrinsicInfor) that is the entropy of distribution of instances into values.

$$GR(X_i) = \frac{IG(X_i)}{IntrinsicInfor(X_i)}. \tag{18}$$

where

$$IntrinsicInfor(X_i) = -\Sigma_{x\in X_i}\frac{n_x}{n} log\frac{n_x}{n} \tag{19}$$

and $n$ is the number of total instances and $n_x$ is the number of instances in the value $x$.

The other approach is to consider the redundancy among features by calculating the score between two features or by using coefficients in the model. Most commonly used methods in are as follows.

7) Minimum Redundancy Maximum Relevance (mRMR): 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. 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 & Peng, 2005).

$$F(X_i, C) = \left[\Sigma_j n_j(\overline{v}_{ij} - \overline{v_i})^2 \middle/ (l-1)\right] \middle/ \sigma^2 \tag{22}$$

where$\overline{v_i}$is the average across all observations in$X_i$, $\overline{v}_{ij}$is the average of$X_i$within the $j$th class ($j = 1,\ldots,l$), and$\sigma^2 = \left[\Sigma_j(n_j - 1)\sigma_j^2\right] \middle/ (n-l)$is the pooled variance ($n_j$and$\sigma_j^2$are the size and the variance of the $j$th class, respectively). For $l = 2$, 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$X_i$and$X_k$, which is denoted by$c(X_i, X_k)$, is chosen. Hence the $F$ test correlation difference

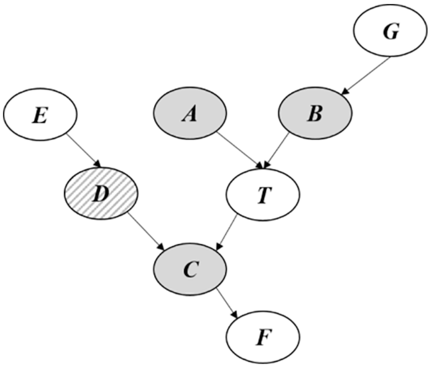

**Fig. 1.** Example of a BN. Nodes (circles): features, edges (arrows): conditional dependencies. Labels and dependencies are explained in the text.

(FCD) and the $F$ test correlation quotient (FCQ) can be defined as follows.

$$FCD = \max_{X_i \in X}\left[F(X_i, C) - \frac{1}{d}\sum_{X_k \in X} c(X_i, X_k)\right] \tag{23}$$

$$FCQ = \max_{X_i \in X}\left[F(X_i, C) \Big/ \frac{1}{d}\sum_{X_k \in X} c(X_i, X_k)\right] \tag{24}$$

The time complexity of mRMR is $O(n \cdot d^2)$.

8) Support Vector Machine Weight Vector (SVM WV) (Guyon et al., 2002): In a SVM, weight vector, $w \in R^d$ can be used to decide the importance of each feature. The larger magnitude of $i$th component of $w$ is, the $i$th feature plays a more important role in the decision function. We thus rank features according to $\{w_i^2\}_{i=1}^d$. This stands on the basis that for the linear kernel, if $w_i^2 = 0$, the optimal hyperplane is the same for the deletion of the $i$th input feature. Even if $w_i^2$ is not zero, if the value is small, the deletion of the $i$th input variable does not affect very much for the optimal hyperplane. Thus, we can delete the feature with the minimum $w_i^2$ (Abe, 2005). The core of an SVM is a quadratic programming problem (QP), separating support vectors from the rest of the training data.

## 3. Theoretical background

### 3.1. Bayesian network (BN) and MB

A BN 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 (Figure 1).

If the Markov condition property holds in a BN, then a node is independent from all nodes other than its descendants when conditioned on its parents (Pearl, 2014). Therefore, a BN consists of a qualitative part in the form of a DAG and a quantitative part in the form of conditional probabilities (Van Harmelen, Lifschitz, & Porter, 2008).

**Definition 1. Faithfulness**

A joint probability distribution $P$ over random variable set $U$ is faithful to a DAG $G$ if and only if all dependencies entailed by $G$ and the Markov condition property are also present in $P$. A data-generating process $K$ is faithful to $G$, if $K$ in the sample limit produces data with joint probability distribution $P$, and $P$ is faithful to $G$. A dataset is faithful to a $G$ if in the sample limit the data was generated by $K$ that is faithful to $G$.

**Definition 2. MB**

Given the faithfulness assumption, from the probability perspective, the MB of a target variable $T$, denoted by $MB(T)$, is a minimal set of variables conditioned on which all other variables $F$ are independent of $T$. In the graphical perspective, $MB(T)$ is the union of parent, child ($PC$), and parent of children (spouse, $SP$) nodes of $T$ (i.e.,$MB(T) = PC(T) \cup SP(T)$). If a dataset from the joint probability distribution $P$ is faithful to DAG $G$, $T$ has a unique MB, $MB(T)$. For example, in Fig. 1, the parent and child nodes of $T$ are $PC(T) = \{A, B, C\}$ and the spouse nodes are$SP(T) = \{D\}$. So, $MB(T) = \{A, B, C, D\}$. It means that node $E$, $F$, and $G$ are independent of $T$ conditioned on $MB(T)$ (Fu & Desmarais, 2010).

The MB has been proven theoretically to be an optimal set of features that does not change the original target distribution (Koller & Sahami, 1996); i.e., $MB(T)$ includes sufficient information to explain $T$ fully. Based on this fact, the MB can be utilized as feature selection.

### 3.2. MB Discovery algorithm

To use the MB as feature selection, we need an algorithm to find the MB of a target variable among features and we call this an MB discovery algorithm. It uses a conditional independence (CI) test to identify CI between features. Therefore, all MB discovery algorithms begin with two basic assumptions: (1) that every CI entailed by a DAG and the Markov condition must be presented in a joint distribution (Fu & Desmarais, 2008, 2010; Pearl, 2014), and (2) that the CI test always gives a correct result. Existing MB discovery algorithms can be divided into two approaches: divide-and-conquer and grow-and-shrink. The former uses topology information for the BN induction and thus requires additional computation burden to discover topology information, while the latter is generally simpler and faster. Therefore, the grow-and-shrink approach is more appropriate than the divide-and-conquer approach as a feature selection method. The most commonly-used grow-and-shrink algorithm is Interleaved Incremental Association MB (Inter-IAMB) (Tsamardinos, Aliferis, Statnikov, & Statnikov, 2003) (**Algorithm 1**) which consists of two iterative steps: adding a candidate feature to the estimated set of MB (lines 1–5) and then immediately eliminating invalid candidates (false positives) caused by the candidate feature (lines 6–10). In the algorithm, $D$, $T$ and $\alpha$ are data, a target variable and a significance level respectively. $Indep(X, Y|Z)$ is the degree (or score) of CI between $X$ and $Y$ given $Z$, which is the $p$-value of a CI test. Note that Inter-IAMB is proven to find MBs correctly.

Algorithm 1. Pseudo code of Inter-IAMB

```
Inter-IAMB($\boldsymbol{D}, \boldsymbol{T}$)
/* add true positives to $MB$ /
1 $MB = \varnothing$
2 repeat
3 $Y = arg : \min_{X \in (U \backslash MB \backslash \{T\})} Indep(T, X|MB)$
4 if $Indep(T, Y|MB) < \alpha$ then
5    $MB = MB \cup \{Y\}$
   /* remove false positives from $MB$ /
6    $rmv = \varnothing$
7      for each $X \in (MB \backslash \{Y\})$ do
8        if $Indep(T, X|MB \backslash \{X\}) \geq \alpha$ then
9          $rmv = rmv \cup \{X\}$
10           $MB = MB \backslash rmv$
12 return $MB$
```

### 3.3. CI test

Identification of the optimal features set entails use of CI tests that are statistical methods to determine that two features $X$ and $Y$ are conditionally independent given the set of features $\boldsymbol{Z}$. For categorical features, an MB discovery algorithm implements a $G^2$ test or a $\chi^2$ test (Pena, Nilsson, Björkegren, & Tegnér, 2007), which are count-based CI tests that use a contingency table. Given a reasonable amount of data, these tests reach the same conclusions. For continuous features, an MB discovery algorithm conducts a Fisher's $z$ test or a Student's $t$ test (Pena et al., 2007), which are model-based CI tests that use linear regression.

Although MB feature selection can be applied in classification and regression by employing different CI tests, it can only deal with data that are of the same type as the target, i.e., classification with categorical features, and regression with continuous features.

### 3.4. Relevance

According to Kohavi and John (1997), a formal definition of relevance is as follows. Define $S$ as the set of all features except $X_i$.

**Definition 3. Strong relevance**

A feature $X_i$ is strongly relevant to $T$ iff there exists an assignment of values $x, t, s$ for which $P(X_i = x, S = s)\rangle 0$ such that

$$P(T = t|X_i = x, S = s) \neq P(T = t|S = s) \tag{25}$$

**Definition 4. Weak relevance**

A feature $X_i$ is weakly relevant to $T$ iff it is not strongly relevant and there exists a subset $Z \subset S$, and an assignment of values $x, t, z$ for which $P(X_i = x, Z = z)\rangle 0$ such that

$$P(T = t|X_i = x, Z = z) \neq P(T = t|Z = z) \tag{26}$$

**Definition 5. Relevance**

A feature is relevant to $T$ if it is weakly or strongly relevant to $T$. A feature is irrelevant to $T$ if it is not relevant to $T$.

The next theorems associate relevance and $MB(T)$.

**Theorem 1.** In a faithful BN, a feature $X_i$ is strongly relevant, iff $X_i \in MB(T)$.

**Theorem 2.** In a faithful BN, a feature $X_i$ is weakly relevant, iff it is not strongly relevant and there is an undirected path from $X_i$ to $T$.

In Fig. 1, $A, B, C$ and D are strongly relevant features to $T$, and $E, F$ and $G$ are weakly relevant features to $T$. In this figure, all features are relevant.

## 4. Proposed method

### 4.1. MB Ranking

We develop a new feature ranking method by embedding the concept of relevance into the MB. First, we suggest the way to measure the quantity of relevance to $T$ based on the MB; a strongly relevant feature $X_i$ in $MB(T)$ can be quantified as strength of dependence to $T$ given all features in $MB(T)$ except for $X_i$, i.e., $dep(T,X_i|MB(T)) \setminus \{X_i\})$. This can be regarded as a relative conditional dependence to $T$ of $X_i$ in $(T)$. Meanwhile, a weakly relevant feature $X_k$ not in $(T)$ can be quantified as strength of dependence to $T$ given $(T)$, i.e., $(T,X_k|\ MB(T))$. This can be regarded as an absolute conditional dependence to $T$ of $X_k$ not in$(T)$. Because of the MB property, the minimum value of strongly relevant features is greater than the maximum value of weakly relevant features.

The method, called MB Ranking (**Algorithm 2**), is simple, which is the strength of this method; it implements a CI test for each feature depending on whether a feature is in $MB(T)$ or not (lines 1–5), and then sorts statistics (or $p$-values) of CI tests in descending order (line 6). In fact, MB Ranking requires $(T)$ as an input; it can be obtained by any MB discovery algorithm. $dep(X, Y|Z)$ is the strength (or degree) of the dependence between$X$and$Y$given $Z$, which is the statistic of a CI test or negative $p$-value of a CI test: the larger the value, the higher relevance. The CI test to be used in this method should be a new CI test because microarray data consist of continuous explanatory variables and categorical target variable; in other words, it has the type-inconsistency problem. Time complexity of this method for calculating the degree of conditional dependency is $O(MB(T) \cdot n \cdot d)$ under the assumption that $MB(T)$ is given. That is, total time complexity should be depending on the time complexity for finding $MB(T)$ (i.e., which MB discovery algorithm is used).

Algorithm 2. Pseudo code of MB Ranking method

```
MB Ranking($D, T, MB(T)$)
1. for each $X = U - \{T\}$
2. if $X \in MB(T)$ then
3.    $score(X) = dep(T, X|MB(T) \setminus \{X\})$
4. else if $X \notin MB(\mathrm{T})$ then
5.    $score(X) = dep(T, X|MB(T))$
6. Sort scorein descending order.
7. return ordered score
```

We used Inter-IAMB to find an MB of the target due to its speed, simplicity, and soundness and named this algorithm Inter-IAMB Ranking (Inter-IAMBR) (**Algorithm 3**). The complexity of Inter-IAMBR is $O(MB(T)^2 \cdot n \cdot d)$. We note that any other MB discover algorithms can be used instead of Inter-IAMB.

Algorithm 3. Pseudo code of Inter-IAMBR algorithm

```
Inter-IAMBR($D, T$)
1. i($T$) = Inter-IAMB($D, T$)
2. $MBRank(T)$ = MB Ranking($D, T, MB(T)$)
3. return $MBRank(T)$
```

### 4.2. CI test for microarray data

As aforementioned, MB feature selection cannot be applied to data with the type inconsistency problem such as classification of microarray data (i.e., classification with continuous features) because CI tests exist only for single type data. Therefore, we suggest the use of generalized CI test, which is based on the likelihood ratio test to analyze this kind of mixed-type data (i.e., the target variable is categorical, but features are continuous variables) (Lee, Jeong, & Jun 2020). The LR test is used to compare the goodness of fit of the null and the alternative models, where the null model is a special case of the other (i.e., nested models). The test is based on the LR Λ, which expresses how many times more likely the data are under one model than the other, and the test statistic is –2logΛ. The test statistic approximately follows a $\chi^2$ distribution with degrees of freedom (df) equal to the difference between the dfs of the alternative model and the null model (Wilks, 1938). Based on this fact, we construct a problem-specific model, in which the likelihood can be obtained, and use the LR test to determine the CI. The problem specific model for classification of Microarray data can be based on a logistic regression model. Given a classification problem with continuous features, we construct a logistic regression model that fits the situation that we are interested in (Eq. 1). Let $Y$ be a target with $l$ categorical values, $X$ be a candidate feature with a continuous value, and $Z$ be a given feature with a continuous value. In other words, researchers would like to know the relationship (i.e., CI) between cancer $Y$ with $l$ classes (for example, the target has different states such as normal, at risk, and cancer or different types of tumor such as Fibroma, Lipoma, and Hemangioma) and gene $X$ given information of gene $Z$. Then, we can construct the following model for $j = 1, \cdots, l-1$

$$\mathrm{logit}[P(Y = j)/P(Y = l)] = \alpha_j + \beta_j X + \gamma_j Z \tag{3}$$

where $\alpha_j$ is an intercept and $\beta_j$, $\gamma_j$ are coefficients of $X$, $Z$. We use this model to obtain the LR and perform the LR test, which determines the CI between $X$ and $Y$ given $Z$ under the null hypothesis $H_0 : \beta_1 = \cdots = \beta_{l-1} = 0$ and df $= l - 1$. If the null hypothesis is true, then $X$ and $Y$ are conditionally independent given $Z$. In other words, $X$ gene cannot provide any further information of $Y$ cancer when the information of $Z$ gene is given. The given feature $Z$ can be extended to a multivariate case $\boldsymbol{Z}$, and the test is proceeded in the same way. The thing that also should be noted is that this new test allows us to handle even multiclass classification problems very smoothly. We embed this new test into the algorithm of MB

**Table 1**
Description of microarray datasets used for classification; number in parenthesis: number of target variables.

| Dataset | # of features | # of observations | # of classes |
|---|---|---|---|
| Leukemia 2 | 7,129 | 72 | 2 |
| Prostate | 12,600 | 102 | 2 |
| SRBCT | 2,308 | 83 | 4 |
| Lung | 12,600 | 203 | 5 |
| Ovarian | 15,114 | 253 | 2 |
| MLL | 12,582 | 72 | 3 |

Ranking.

## 5. Experimental results and discussion

### 5.1. Microarray data

Microarray data are a large source of genetic data, which, upon proper analysis, could enhance our understanding of biology and medicine. Many microarray experiments have been designed to investigate the genetic mechanisms of cancer, and analytical approaches have been applied to classify different types of cancer or distinguish between contaminated and non-contaminated tissue. In the last ten years, machine learning methods have been investigated in microarray data analysis. Several approaches have been tried in order to (i) distinguish between contaminated and non-contaminated samples, (ii) classify different types of cancer, and (iii) identify subtypes of cancer that may progress aggressively. All these investigations are seeking to generate biologically meaningful interpretations of complex datasets that are sufficiently interesting to drive follow-up experimentation (Hira & Gillies, 2015). We extracted six datasets from the UCI repository of Kent Ridge Bio-medical Dataset (http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html) (Table 1). Note that some datasets are multiclass problems.

**Table 2**
LOOCV classification error rate with SVM of six gene expression datasets for different gene selection methods using 10 to 130 selected genes at 20 intervals. Bold: minimum error rate at each number of selected features in each dataset; red bold: grand minimum error in each dataset. Single asterisk indicates $P < 0.05$ by $z$ test; double asterisk, $P < 0.01$.

| Dataset | Method | SVM | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 10 | 30 | 50 | 70 | 90 | 110 | 130 |
| Leukemia 2 | Inter-IAMBR | **2.78** | **2.78** | 5.56 | 5.56 | 5.56 | **2.78** | **1.39** |
| | ERGS | 11.11 | 9.72 | 5.56 | 6.94 | 6.94 | **2.78** | 4.17 |
| | IFSER | 9.72 | 9.72 | 9.72 | 6.94 | 6.94 | 6.94 | 8.33 |
| | PCC | 19.44 | 8.33 | 8.33 | 6.94 | 8.33 | 9.72 | 15.28 |
| | Relief-F | 12.50 | 6.94 | 4.17 | **2.78** | 4.17 | **2.78** | 2.78 |
| | IG | 4.17 | 5.56 | 9.72 | 4.17 | **2.78** | **2.78** | 5.56 |
| | GR | 6.94 | 4.17 | **2.78** | 4.17 | 5.56 | 5.56 | 5.56 |
| | SVM WV | 5.56 | 6.94 | 4.17 | 4.17 | 4.17 | 4.17 | 4.17 |
| Prostate | Inter-IAMBR | **0.98**** | **1.96**** | **3.92** | 2.94 | **2.94** | **3.92** | 5.88 |
| | ERGS | 4.90 | 13.73 | 9.80 | 9.80 | 8.82 | 8.82 | 7.84 |
| | IFSER | 15.69 | 11.76 | 8.82 | 7.84 | 6.86 | 9.80 | 8.82 |
| | PCC | 10.78 | 14.71 | 13.73 | 13.73 | 12.75 | 9.80 | 12.75 |
| | Relief-F | 9.80 | 10.78 | **3.92** | 4.90 | 4.90 | 6.86 | 5.88 |
| | IG | 10.78 | 8.82 | 8.82 | 6.86 | 10.78 | 11.76 | 8.82 |
| | GR | 8.82 | 10.78 | 10.78 | 11.76 | 11.76 | 10.78 | 7.84 |
| | SVM WV | 6.86 | 7.84 | 5.88 | **0.98** | **2.94** | 4.90 | **2.94** |
| SRBCT | Inter-IAMBR | **0.00*** | **1.20** | 1.20 | 1.20 | 1.20 | 1.20 | 1.20 |
| | ERGS | 19.28 | 10.84 | 9.64 | 9.64 | 8.43 | 7.23 | 6.02 |
| | IFSER | 34.94 | 27.71 | 20.48 | 18.07 | 12.05 | 12.05 | 14.46 |
| | PCC | 9.64 | **1.20** | 1.20 | 1.20 | 1.20 | 0.00 | 1.20 |
| | Relief-F | 4.82 | 2.41 | 1.20 | **0.00** | **0.00** | **0.00** | **0.00** |
| | IG | 2.41 | **1.20** | 1.20 | **0.00** | **0.00** | **0.00** | **0.00** |
| | GR | 13.25 | 2.41 | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** |
| | SVM WV | 6.02 | 3.61 | 2.41 | 2.41 | 1.20 | 1.20 | **0.00** |
| Lung | Inter-IAMBR | **5.42*** | 5.42 | 6.40 | 4.93 | 4.43 | **3.45** | 5.42 |
| | ERGS | 23.15 | 12.81 | 8.37 | 5.91 | 6.40 | 4.93 | 5.42 |
| | IFSER | 11.33 | 8.87 | 8.37 | 7.39 | 7.88 | 7.39 | 7.39 |
| | PCC | 14.78 | 11.82 | 14.78 | 10.34 | 9.36 | 8.87 | 8.37 |
| | Relief-F | 9.36 | 8.87 | 9.36 | 9.85 | 10.84 | 7.88 | 8.87 |
| | IG | 12.81 | 6.90 | 9.85 | 5.91 | 6.40 | 6.40 | 5.91 |
| | GR | 23.65 | 20.20 | 6.40 | 8.37 | 7.88 | 5.91 | 6.90 |
| | SVM WV | 9.36 | **4.43** | **5.42** | **3.94** | **3.94** | 4.43 | **3.94** |
| Ovarian | Inter-IAMBR | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** |
| | ERGS | 7.91 | 0.40 | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** |
| | IFSER | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** |
| | PCC | 7.91 | 0.40 | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** |
| | Relief-F | 7.51 | 3.56 | 1.98 | 1.98 | 2.77 | 3.95 | 2.77 |
| | IG | 1.58 | 0.40 | 0.40 | 0.40 | **0.00** | **0.00** | **0.00** |
| | GR | 1.98 | 1.58 | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** |
| | SVM WV | 1.98 | 0.40 | 0.40 | **0.00** | **0.00** | **0.00** | **0.00** |
| MLL | Inter-IAMBR | **2.78** | **2.78** | 5.56 | 5.56 | 5.56 | **2.78** | **1.39** |
| | ERGS | 33.33 | 29.17 | 29.17 | 18.06 | 15.28 | 13.89 | 12.50 |
| | IFSER | 43.06 | 27.78 | 22.22 | 12.50 | 12.50 | 13.89 | 15.28 |
| | PCC | 19.44 | 8.33 | 8.33 | 6.94 | 8.33 | 9.72 | 15.28 |
| | Relief-F | 12.50 | 6.94 | 4.17 | **2.78** | 4.17 | **2.78** | 2.78 |
| | IG | 4.17 | 5.56 | 9.72 | 4.17 | **2.78** | **2.78** | 5.56 |
| | GR | 6.94 | 4.17 | **2.78** | 4.17 | 5.56 | 5.56 | 5.56 |
| | SVM WV | 5.56 | 6.94 | 4.17 | 4.17 | 4.17 | 4.17 | 4.17 |

**Table 3**
LOOCV classification error rate with *k*-NN of six gene expression datasets for different gene selection methods using 10 to 130 selected genes at 20 intervals. Bold: minimum error rate at each number of selected features in each dataset; red bold: grand minimum error in each dataset. Single asterisk indicates $P < 0.05$ by $z$ test; double asterisk, $P < 0.01$.

| Dataset | Method | *k*-NN | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 10 | 30 | 50 | 70 | 90 | 110 | 130 |
| Leukemia 2 | Inter-IAMBR | 9.72 | 15.28 | **2.78** | **5.56** | 5.56 | **2.78** | **1.39** |
| | ERGS | 8.33 | 8.33 | 9.72 | 8.33 | 8.33 | 6.94 | 2.78 |
| | IFSER | 8.33 | 9.72 | 6.94 | 8.33 | 6.94 | 5.56 | 6.94 |
| | PCC | 15.28 | 11.11 | 11.11 | 19.44 | 18.06 | 13.89 | 11.11 |
| | Relief-F | 6.94 | 5.56 | 8.33 | 6.94 | 6.94 | 6.94 | 5.56 |
| | IG | 8.33 | 6.94 | 9.72 | 8.33 | 9.72 | 6.94 | 5.56 |
| | GR | 9.72 | 8.33 | 5.56 | 6.94 | 6.94 | 6.94 | 6.94 |
| | SVM WV | **5.56** | **5.56** | 4.17 | **5.56** | **4.17** | 4.17 | 4.17 |
| Prostate | Inter-IAMBR | 10.78 | 11.76 | 9.80 | **4.90** | **6.86** | **3.92*** | **3.92**** |
| | ERGS | **5.88** | 8.82 | 9.80 | 8.82 | 9.80 | 8.82 | 9.80 |
| | IFSER | 7.84 | 7.84 | 9.80 | 10.78 | 9.80 | 10.78 | 10.78 |
| | PCC | 13.73 | 18.63 | 28.43 | 23.53 | 24.51 | 24.51 | 29.41 |
| | Relief-F | **5.88** | **6.86** | 9.80 | 8.82 | 9.80 | 7.84 | 7.84 |
| | IG | 6.86 | 7.84 | **6.86** | 6.86 | 9.80 | 9.80 | 9.80 |
| | GR | 6.86 | 9.80 | 7.84 | 8.82 | 9.80 | 13.73 | 11.76 |
| | SVM WV | 8.82 | **6.86** | 7.84 | 7.84 | 11.76 | 11.76 | 11.76 |
| SRBCT | Inter-IAMBR | **0.00**** | 3.61 | 3.61 | 4.82 | 6.02 | 6.02 | 4.82 |
| | ERGS | 26.51 | 10.84 | 12.05 | 10.84 | 10.84 | 12.05 | 10.84 |
| | IFSER | 28.92 | 19.28 | 14.46 | 13.25 | 19.28 | 18.07 | 18.07 |
| | PCC | 16.87 | 18.07 | 14.46 | 14.46 | 13.25 | 9.64 | 7.23 |
| | Relief-F | 3.61 | **0.00** | 1.20 | **0.00** | **0.00** | **0.00** | 1.20 |
| | IG | 9.64 | 4.82 | 1.20 | 1.20 | 1.20 | **0.00** | 2.41 |
| | GR | 25.30 | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | 1.20 |
| | SVM WV | 12.05 | 6.02 | 6.02 | 4.82 | 4.82 | **0.00** | **0.00** |
| Lung | Inter-IAMBR | **7.88** | **7.88** | **5.42** | **4.43** | **3.94** | 6.40 | **4.93** |
| | ERGS | 23.15 | 8.87 | 6.90 | 6.40 | 5.91 | 5.91 | 5.91 |
| | IFSER | 24.63 | 11.33 | 8.87 | 8.87 | 6.90 | **5.42** | 6.40 |
| | PCC | 20.20 | 16.26 | 14.29 | 10.84 | 15.76 | 13.30 | 11.82 |
| | Relief-F | 11.82 | 11.82 | 11.33 | 11.82 | 12.32 | 10.84 | 10.84 |
| | IG | 16.26 | 10.34 | 10.34 | 10.84 | 10.34 | 6.90 | 5.42 |
| | GR | 26.11 | 23.15 | 6.40 | 8.37 | 6.90 | 7.88 | 8.87 |
| | SVM WV | 9.36 | 8.37 | 5.91 | 5.91 | 5.42 | 5.91 | 5.91 |
| Ovarian | Inter-IAMBR | **0.00**** | 0.40 | 0.79 | 1.19 | 1.58 | 1.58 | 1.58 |
| | ERGS | 18.18 | 4.35 | **0.00** | **0.00** | **0.00** | 0.40 | 0.40 |
| | IFSER | 18.18 | 4.35 | **0.00** | **0.00** | 0.40 | 0.40 | 0.40 |
| | PCC | 15.02 | 12.65 | 11.46 | 13.83 | 14.23 | 17.00 | 17.39 |
| | Relief-F | 1.98 | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** | 0.40 |
| | IG | 1.98 | 1.19 | **0.00** | **0.00** | 0.40 | 0.79 | 0.79 |
| | GR | 1.98 | **0.00** | 0.40 | 0.79 | 0.40 | 1.19 | 0.79 |
| | SVM WV | 3.16 | **0.00** | 0.40 | 0.40 | 0.40 | **0.00** | **0.00** |
| MLL | Inter-IAMBR | 9.72 | 15.28 | **2.78** | **5.56** | 5.56 | **2.78** | **1.39** |
| | ERGS | 33.33 | 27.78 | 29.17 | 15.28 | 15.28 | 16.67 | 13.89 |
| | IFSER | 34.72 | 18.06 | 15.28 | 9.72 | 9.72 | 8.33 | 12.50 |
| | PCC | 15.28 | 11.11 | 11.11 | 19.44 | 18.06 | 13.89 | 11.11 |
| | Relief-F | 6.94 | **5.56** | 8.33 | 6.94 | 6.94 | 6.94 | 5.56 |
| | IG | 8.33 | 6.94 | 9.72 | 8.33 | 9.72 | 6.94 | 5.56 |
| | GR | 9.72 | 8.33 | 5.56 | 6.94 | 6.94 | 6.94 | 6.94 |
| | SVM WV | **5.56** | **5.56** | 5.56 | **5.56** | **4.17** | 4.17 | 4.17 |

### 5.2. Comparative evaluation

We investigate the performance of the proposed algorithm for microarray data with results achieved by other feature ranking methods. Note that there are no wrapper or embedded related methods in this experiment. We selected seven gene selection methods to be compared with Inter-IAMB Ranking: ERGS, IFSER, PCC, Relief-F, IG, GR, and SVM WV. These are most commonly used ranking methods for gene selection and proven to be well-performed in previous studies (Guyon et al., 2002; Van't Veer et al., 2002; Chandra & Gupta, 2011; Wang et al., 2014). mRMR was excluded from this comparison because it is not feasible in these high-dimensional datasets due to its complexity proportional to the square of the number of features. Two popular classifiers, SVM (Cortes & Vapnik, 1995) and *k*-NN (Fix & Hodges, 1989) (one parametric method and one non-parametric method for avoiding classifier bias) were used to assess the error rate of selected features by using leave-one-out cross validation (LOOCV), which is a proper strategy to give a relatively comprehensive comparison on the performances in high-dimensional data. The seven gene subsets of top 10 to 130 at 20 intervals are selected to highlight the effectiveness of the proposed method for all datasets (Tables 2 and 3). We used the $z$ test of the independent sample approach for LOOCV to compare top two algorithms (Wong, 2015) at every experiment; single asterisk and double asterisk respectively indicate $P < 0.1$ and $P < 0.05$ by the $z$ test. The experiments were conducted by MATLAB on an Intel® Core$^{TM}$ i7-7500U CPU operating at 2.90 GHz, with 16.0 GB of RAM.

Most notably, Inter-IAMBR outperformed other methods for all combinations of datasets and classifiers given. Furthermore, the significant gap between top two gene selection methods is observed only in the proposed method: three datasets with SVM and two datasets with *k*-NN; this is a remarkable result considering LOOCV and a small number of samples. To understand the superior performance of the proposed method, we discuss about the aforementioned gene selection methods, including the proposed method, in terms of three aspects: whether the method is univariate or multivariate, how each method handles type-inconsistent data (i.e., continuous features and nominal target), and

**Table 4**
Comparison among gene selection methods.

| Method | Uni/ multivariate | Type inconsistency solution | Computational complexity |
|---|---|---|---|
| Inter-IAMBR | Multivariate | Generalized CI test | $O(\|MB(T)\|^2 \cdot n \cdot d)$ |
| ERGS | Univariate | Class-based effective range | $O(n \cdot d)$ |
| IFSER | Univariate | Class-based effective range | $O(n \cdot d)$ |
| PCC | Univariate | Regarding nominal as continuous | $O(n \cdot d)$ |
| Relief-F | Univariate | Class-based sample | $O(m \cdot n \cdot d)$ |
| IG | Univariate | Discretization | $O(n \cdot d)$ |
| GR | Univariate | Discretization | $O(n \cdot d)$ |
| SVM WV | Multivariate | Multiclass SVM | $O(n^2 \cdot d) \sim O(n^3 \cdot d)$ |

**Table 5**
Number of total features vs. number of selected features by Inter-IAMB.

| Dataset | Leukemia 2 | Prostate | SRBCT | Lung | Ovarian | MLL |
|---|---|---|---|---|---|---|
| $d$ | 7,129 | 12,600 | 2,308 | 12,600 | 15,114 | 12,582 |
| $\|MB(T)\|$ | 4 | 5 | 4 | 7 | 2 | 4 |

computational complexity (Table 4).

At first, the factor that most affects accuracy comes from the difference between multivariate and univariate methods. Inter-IAMBR is a multivariate method able to consider the interaction among other features as well as the relevance to a target, and thus univariate methods considering the relevance to a target only such as ERGS, IFSER, PCC, Relief-F, IG, and GR might not beat Inter-IAMBR. This can be also seen that SVM WV was well performed in overall. Secondly, data efficiency is an important factor to determine the quality of selected features in high-dimensional data. Inter-IAMBR can be a data-efficient multivariate method because it only utilizes maximum $|\mathrm{CMB}|$ parameters compared to SVM WV that considers $d$ parameters within finite small samples. In fact, $|\mathrm{CMB}|$ can be empirically inferred to be much smaller than $d$ (i.e., $|\mathrm{CMB}| \ll d$) from the result of $|\mathrm{MB}(T)|$ (Table 5).

In further analysis, we analyze the result from other methods. Note that we found that our results show a similar pattern with the previous works with comparison results, such as SVM WV proposal (Guyon et al., 2002), ERGS proposal (Chandra & Gupta, 2011), and IFSER proposal (Wang et al., 2014). To be specific, PCC tends to be inferior to IG and GR. This is because PCC is under the linear assumption, and PCC regards a target as a continuous variable, which would be a serious mistake. To keep the consistency of variable types in classification, discretizing explanatory variables is more desirable than treating the target as a continuous one, but discretization-based methods may obtain different results depending on the way to discretize. In addition, these two kinds of methods cannot avoid the information loss implicitly. Secondly, IFSER, ERGS and Relief-F are developed to solve classification directly without loss of information. IFSER and ERGS provided unstable and even relatively poor error rates; the effectiveness of effective range can vary depending on the dataset. In contrast, Relief-F achieved stable results for several datasets even though it is not optimal. However, it has several factors to be controlled by a user such as the number of neighborhoods (only the nearest neighborhood is considered in this study), the number of updating operations and the way to sample, which can affect the result. For above issues, Inter-IAMBR has relative advantages. It solves the type inconsistency problem of microarray data by embedding the generalized CI test. In addition, only significant level should be considered before between reference values 0.01 or 0.05. The weakest point of Inter-IAMBR compared to other methods is the time complexity. The average time complexity is summarized as $O(|MB(T)|^2 \cdot n \cdot d)$, but the detailed average time complexity is $O(|MB(T)|^2 \cdot n \cdot d + \left|MB(T)\right| \cdot n \cdot d + d\log d)$. Each term corresponds to the produced time complexity for MB discovery, relevance calculation and ranking in order. However, this is reasonable for a multivariate method because it does not expand to the power of dimensionality.

## 6. Conclusions

In this study, we proposed an efficient multivariate feature ranking method for classification of high-dimensional microarray data. We combined the relevance concept with the MB and designed a new CI test able to deal with the type inconsistency problem of microarray data. The proposed method, called MB Ranking, measures the relative relevance of each gene to a target by considering the redundancy among other genes within a reasonable computational complexity. As expected, MB-based feature ranking outperformed commonly used-univariate ranking methods in the experiment with six different microarray datasets. It also yields better results even compared with the other multivariate feature ranking method due to the advantage of data efficiency. The new method can be regarded as an extension of existing MB feature selection, which enables to provide the ranking of features as well as subset of features and to handle the type-inconsistency of the data. We believe that the new method can be a beneficial alternative in classification analysis of gene expression data compared to other methods in terms of prediction performance, simplicity of implementation, and selectability of output form.

However, it is worth pointing out several limitations and future directions for the improvement of the work and the proposed method. First of all, in this study, the methods compared with the proposed method were limited to well-known filter methods. With computing power continuing to increase exponentially, time complexity may no longer be a burden. We will further investigate the performance of the proposed method compared with state-of-the-art methods including those in the wrapper and embedded families. Secondly, our research focused on prediction accuracy analysis, not covering the investigation of top-ranked features (i.e., interpretability of important features). We plan to apply this method to validate/discover well-known/unknown biomarkers on new real-world microarray data on cancer beyond benchmark datasets. Finally, In future work, we may consider changing Inter-IAMB to the latest alternatives such as STMB (Gao & Ji, 2016) and BAMB (Ling et al., 2019) to improve data efficiency in the process of finding MB, which might provide better results for gene selection.

### CRediT authorship contribution statement

**Junghye Lee:** Conceptualization, Methodology, Software. **In young Choi:** Writing - review & editing. **Chi-Hyuck Jun:** Supervision, Writing - review & editing.

### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgement

The work of J. Lee and I.Y. Choi was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MSIT) (No 2018R1C1B5086611, 2019R1A5A2027588 and 2020R1C1C1011063). The work by C.-H. Jun was supported by NRF grant funded by the Korean government (MEST) (No 2020R1A2C1005442). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-01336, Artificial Intelligence graduate school support (UNIST)) and the 2020

Research Fund (1.200098.01) of UNIST.

## References

Abe, S. (2005). Support vector machines for pattern classification (Vol. 2, p. 44). London: Springer.

Abdulqader, D. M., Abdulazeez, A. M., & Zeebaree, D. Q. (2020). Machine Learning Supervised Algorithms of Gene Selection: A Review. *Machine Learning, 62*(03).

Almugren, N., & Alshamlan, H. (2019). A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE Access, 7*, 78533–78548.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences, 96*(12), 6745–6750.

Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. A. (2015). Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 13*(5), 971–989.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks, 5*(4), 537–550.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., & Yakhini, Z. (2000). Tissue classification with gene expression profiles. *Journal of Computational Biology, 7*(3–4), 559–583.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2015). *Feature Selection for High-Dimensional Data*. London: Springer.

Brown, G. (2009). A new perspective for information theoretic feature selection. In Artificial intelligence and statistics (pp. 49-56).

Chandra, B., & Gupta, M. (2011). An efficient statistical feature selection approach for classification of gene expression data. *Journal of Biomedical Informatics, 44*(4), 529–535.

Chen, H., Zhang, Y., & Gutman, I. (2016). A kernel-based clustering method for gene selection with gene expression data. *Journal of Biomedical Informatics, 62*, 12–20.

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

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 215–242.

Devi Arockia Vanitha, D. A., Devaraj, D., & Venkatesuluc, M. (2015). Gene expression data classification using support vector machine and mutual information-based gene selection. *Procedia Computer Science, 47*, 13–21.

Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology, 3*(02), 185–205.

Duch, W., Wieczorek, T., Biesiada, J., & Blachnik, M. (2004). Comparison of feature ranking methods based on information entropy. In 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541) (Vol. 2, pp. 1415-1419). IEEE.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*(2), 179–188.

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique, 57*(3), 238–247.

Fu, S., & Desmarais, M. C. (2008). In *Fast Markov blanket discovery algorithm via local learning within single pass* (pp. 96–107). Berlin, Heidelberg: Springer.

Fu, S., & Desmarais, M. C. (2010). Markov blanket based feature selection: a review of past decade. In Proceedings of the world congress on engineering (Vol. 1, pp. 321-328). Newswood Ltd.

Jain, L. C. (Ed.). (2008). *Computational intelligence: a compendium* (Vol. 21). Warsaw, Poland: Springer.

Gao, T., & Ji, Q. (2016). Efficient Markov blanket discovery and its application. *IEEE transactions on Cybernetics, 47*(5), 1169–1179.

García, V., & Sánchez, J. S. (2015). Mapping microarray gene expression data into dissimilarity spaces for tumor classification. *Information Sciences, 294*, 362–375.

García, V., Sánchez, J. S., Cleofas-Sánchez, L., Ochoa-Domínguez, H. J., & López-Orozco, F. (2017). An insight on the 'large G, small n'problem in gene-expression microarray classification. In Iberian Conference on Pattern Recognition and Image Analysis (pp. 483-490). Springer, Cham.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., & Bloomfield, C. D. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *science, 286*(5439), 531–537.

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

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning, 46*(1–3), 389–422.

Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics, 2015.

Hoque, N., Ahmed, H. A., Bhattacharyya, D. K., & Kalita, J. K. (2016). A fuzzy mutual information-based feature selection method for classification. *Fuzzy Information and Engineering, 8*(3), 355–384.

Hsu, H. H., Hsieh, C. W., & Lu, M. D. (2011). Hybrid feature selection by combining filters and wrappers. *Expert Systems with Applications, 38*(7), 8144–8150.

Ke, W., Wu, C., Wu, Y., & Xiong, N. N. (2018). A new filter feature selection based on criteria fusion for gene microarray data. *IEEE Access, 6*, 61065–61076.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1–2), 273–324.

Koller, D., & Sahami, M. (1996). *Toward optimal feature selection*. Stanford InfoLab.

Kononenko, I. (1994). In *Estimating attributes: analysis and extensions of RELIEF* (pp. 171–182). Berlin, Heidelberg: Springer.

Lamba, M., Munjal, G., & Gigras, Y. (2018). Feature Selection of Micro-array expression data (FSM)-A Review. *Procedia Computer science, 132*, 1619–1625.

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., … Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9*(4), 1106–1119.

Lê Cao, K. A., Bonnet, A., & Gadat, S. (2009). Multiclass classification and gene selection with a stochastic algorithm. *Computational Statistics & Data Analysis, 53*(10), 3601–3615.

Lee, J., Jeong, J. Y., & Jun, C. H. (2020). Markov Blanket-based Universal Feature Selection for Classification and Regression of Mixed-Type Data. *Expert Systems with Applications, 113398*.

Liao, B., Jiang, Y., Liang, W., Peng, L., Peng, L., Hanyurwimfura, D., Li, Z., & Chen, M. (2015). On efficient feature ranking methods for high-throughput data analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 12*(6), 1374–1384.

Ling, Z., Yu, K., Wang, H., Liu, L., Ding, W., & Wu, X. (2019). Bamb: A balanced Markov blanket discovery approach to feature selection. *ACM Transactions on Intelligent Systems and Technology (TIST), 10*(5), 1–25.

Liu, H., Li, J., & Wong, L. (2002). A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome informatics, 13*, 51–60.

Liu, H., Motoda, H., & Dash, M. (1998). In *A monotonic measure for optimal feature selection* (pp. 101–106). Berlin, Heidelberg: Springer.

Liu, H., & Setiono, R. (1995). In *Chi2: Feature selection and discretization of numeric attributes* (pp. 388–391). IEEE.

Mabu, A. M., Prasad, R., & Yadav, R. (2020). Mining gene expression data using data mining techniques: A critical review. *Journal of Information and Optimization Sciences, 41*(3), 723–742.

Manikandan, G., & Abirami, S. (2018). *A survey on feature selection and extraction techniques for high-dimensional microarray datasets*. Springer, Singapore: In Knowledge Computing and its Applications.

Quinlan, J. R. (2014). *programs for machine learning, C4*, 5.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Elsevier.

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

Peng, H., Long, F., & Ding, C. (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.

Raweh, A. A., Nassef, M., & Badr, A. (2018). A hybridized feature selection and extraction approach for enhancing cancer prediction based on DNA methylation. *IEEE Access, 6*, 15212–15223.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., & Pergamenschikov, A. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics, 24*(3), 227.

Ruiz, R., Riquelme, J. C., & Aguilar-Ruiz, J. S. (2006). Incremental wrapper-based gene selection from microarray data for cancer classification. *Pattern Recognition, 39*(12), 2383–2392.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics, 23*(19), 2507–2517.

Shen, Q., Diao, R., & Su, P. (2012). Feature Selection Ensemble. *Turing-100, 10*, 289–306.

Sun, L., Zhang, X. Y., Qian, Y. H., Xu, J. C., Zhang, S. G., & Tian, Y. (2019). Joint neighborhood entropy-based gene selection method with fisher score for tumor classification. *Applied Intelligence, 49*(4), 1245–1259.

Tang, J., Alelyani, S., & Liu, H. (2014). *Feature selection for classification: A review* (p. 37). Data classification: Algorithms and applications.

Trunk, G. V. (1979). A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 3*, 306–307.

Tsamardinos, I., Aliferis, C. F., Statnikov, A. R., & Statnikov, E. (2003). Algorithms for Large Scale Markov Blanket Discovery. *FLAIRS Conference, 2*, 376–380.

Van Harmelen, F., Lifschitz, V., & Porter, B. (Eds.). (2008). *Handbook of knowledge representation* (Vol. 1). Elsevier.

Wang, A., An, N., Yang, J., Chen, G., Li, L., & Alterovitz, G. (2017). Wrapper-based gene selection with Markov blanket. *Computers in biology and medicine, 81*, 11–23.

Wang, J., Zhou, S., Yi, Y., & Kong, J. (2014). An improved feature selection based on effective range for classification. The Scientific World Journal, 2014.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics, 9*(1), 60–62.

Vanjimalar, S., Ramyachitra, D., & Manikandan, P. (2018). In *A Review on Feature Selection Techniques for Gene Expression Data* (pp. 1–4). IEEE.

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., Van Der Kooy, K., Marton, M.J., Witteveen, A.T., & Schreiber, G. J. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415(6871), 530.