Talanta 85 (2011) 1689–1694

Contents lists available at ScienceDirect

Talanta

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

Unimodal transform of variables selected by interval segmentation purity for classification tree modeling of high-dimensional microarray data

Wen Du, Ting Gu, Li-Juan Tang*, Jian-Hui Jiang, Hai-Long Wu, Guo-Li Shen, Ru-Qin Yu*

State Key Laboratory of Chemo/Biosensing and Chemometrics, Laboratory of Tobacco Chemistry, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, PR China

ARTICLE INFO

Article history:
Received 6 May 2011
Received in revised form 28 June 2011
Accepted 30 June 2011
Available online 7 July 2011

Keywords:
Classification and regression tree
Interval segmentation purity
Variable selection
Gene expression
Mean shift

ABSTRACT

As a greedy search algorithm, classification and regression tree (CART) is easily relapsing into overfitting while modeling microarray gene expression data. A straightforward solution is to filter irrelevant genes via identifying significant ones. Considering some significant genes with multi-modal expression patterns exhibiting systematic difference in within-class samples are difficult to be identified by existing methods, a strategy that unimodal transform of variables selected by interval segmentation purity (UTISP) for CART modeling is proposed. First, significant genes exhibiting varied expression patterns can be properly identified by a variable selection method based on interval segmentation purity. Then, unimodal transform is implemented to offer unimodal featured variables for CART modeling via feature extraction. Because significant genes with complex expression patterns can be properly identified and unimodal feature extracted in advance, this developed strategy potentially improves the performance of CART in combating overfitting or underfitting while modeling microarray data. The developed strategy is demonstrated using two microarray data sets. The results reveal that UTISP-based CART provides superior performance to *k*-nearest neighbors or CARTs coupled with other gene identifying strategies, indicating UTISP-based CART holds great promise for microarray data analysis.

1. Introduction

Microarray based gene expression signatures have shown the potential to be powerful tools for patient stratification, diagnosis of disease, prognosis of survival, assessment of risk group and selection of treatment [1–4]. These signatures are computational rules for deriving a diagnosis from patient's expression profile. Although microarray based molecular diagnosis constitutes a promising approach, so far it is difficult for most machine learning methods to exploit its potential. For typical microarray data, the number of genes (variables) may be times that of samples, making variable redundancy to be a serious problem when constructing a classification model.

To be a good choice, classification and regression trees (CARTs) [5] are frequently used to generate decision rules for microarray data [6–8]. CART is a nonlinear, parameter-free tree-building technique, quite flexible in describing the relationship between independent and dependent variables. Contrast to most of modeling methods, CART has several advantages: (1) the model building by CART is intuitive and comprehensive; (2) being an excellent tool in fitting nonlinear data; (3) suitability for high-dimension data sets, capable automatically select parts of the important variables; (4) free from the singular values, collinearity and variance heterogeneity, etc. However, being a greedy search algorithm, CART is easily relapsing into overfitting. When dealing with microarray gene expression data that frequently full of noises and redundant information originated from a large number of irrelevant genes, exhaustively searching all variables would aggravate the overfitting problem of CART. A straightforward approach is to filter irrelevant genes by identifying significant ones before constructing a CART model.

A gene related to certain disease possibly display systematic expression differences in within-class samples, known as multi-modal expression pattern [9–12]. Such phenomenon is derived from the diversity of pathogenic mechanisms. Properly identifying these multi-modal expressed disease-related genes is much important for building desirable pattern recognition models. It is also crucial for the studies of molecular diagnosis and personalized medicine. However, few existing methods can identify multi-modal expressed genes well due to ignoring the systematic expression differences of a gene in within-class samples. In our previous study, a technique for gene identification based on interval segmentation purity (ISP) has been proposed and demonstrated [11]. Intuitively, a gene discriminative for a class of interest implies that its expression level for the class is significantly different from those of the others, that is, its expression levels for the class are centralized at one or

* Corresponding authors. Tel.: +86 731 88822577; fax: +86 731 88822782.
E-mail addresses: tanglijuan@hnu.edu.cn (L.-J. Tang), rqyu@hnu.cn (R.-Q. Yu).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.06.076

several intervals where the expression levels for the other classes rarely appear. In other words, there are some intervals where the samples from the class of interest are centralized and few other samples present. Then, these intervals can be regarded to have adequate "purity" of samples belonging to the class of interest with respect to other classes. In this sense, the discriminative power of a gene for one class can be measured by the "purity" of the intervals where the expression levels of the gene for the class are centralized. Hence, in the ISP-based gene identification technique ISP was defined as the purity of samples belonging to a certain class in intervals segmented by a mode search algorithm, mean shift [13]. This technique was demonstrated to furnish an advantage over other methods in the capability of selecting discriminative genes, whatever the genes are up-regulated, down-regulated or multi-modal.

Based on the ISP-based gene identification technique [11], herein, we developed a flexible variable treating strategy for CART modeling microarray data, unimodal transform of the variables identified by ISP (UTISP). By invoking the ISP-based technique, significant genes with varied expression patterns are first properly identified. However, it is not so advisable to build a CART model for microarray data straight utilizing the significant genes identified by using ISP-based technique. As commonly plying a binary recursive partitioning strategy, it is difficult for CARTs to generate proper decision rules for multi-modal expressed genes, which may increase the overfitting or underfitting risk of the models. The strategy that unimodal transform of the significant genes is then developed to extract features of significant genes via mapping them into their higher dimensional spaces where unimodal expression patterns of these genes can be disclosed. Unimodal featured variables resulting from the feature extraction procedure offers a flexible variable treating approach for CART while modeling microarray data, thus this strategy can significantly improve the performance of CART in generating classification rules for microarray data. The proposed UTISP-based CART methodology has been evaluated in cancer classification using two public microarray data sets, leukemia data [14] and small, round blue cell tumors data (SRBCTs) [15].

2. Methods

2.1. ISP-based gene identification technique

The idea of the ISP-based gene identifying technique is to segment each variable (gene) axis into intervals or clusters where the expression level values are centralized, followed by the evaluation of the purity of these clusters for a class of interest. The ISP-based technique utilizes a mode search algorithm, mean shift [13,16,17], to disclose the natural clusters of the data, by which the clusters are defined as the unimodal distribution of the probability density function [11]. Mean shift is a computationally efficient algorithm for mode search based clustering. This procedure is essentially a gradient ascent algorithm started from every data points. Then, all data points located in the same unimodal intervals, i.e. in the same clusters, converge at the same modes within a specified error level δ (a parameter to be set in the algorithm, in the present study $\delta = 10^{-4}$), so the cluster membership of a data point can be defined by the mode at which it converges. Suppose that a microarray data set has P genes and N samples with the entries x_{pn} ($p = 1, \ldots, P$; $n = 1, \ldots, N$) representing the expression level of the nth sample, $\mathbf{x}_n$, on the pth gene. Let K be a flat kernel that is the characteristic function with a certain analysis bandwidth h,

$$K(x) = \begin{cases} 1 & \text{if} \quad \|x\| \le h \\ 0 & \text{if} \quad \|x\| > h \end{cases} \tag{1}$$

The mean shift vector in a simple one-dimensional case (on one gene, i.e. on gene p) can be expressed as

$$m(x) = \frac{\sum_{n=1}^{N} x_{pn} K(x - x_{pn})}{\sum_{n=1}^{N} K(x - x_{pn})} - x \tag{2}$$

where x is a data point in a one-dimensional data space, starting from an arbitrary data point, x_{pn}, for $n = 1, \ldots, N$. The analysis bandwidth h is a positive value which can be determined by sensitivity analysis. When $m(x)$ is applied to the original point, x, it results in a new position, x^s. This process can be repeated until $|m(x)| < \delta$ and an iterative procedure is defined in this way:

$$x^{s+1} = x^s + m(x^s) \tag{3}$$

For a kernel with a monotonically decreasing profile, convergence of x^s ($s = 1, \ldots, S$) can be proven. The iterative mean shift procedure is, in essence, a gradient ascent method where the step size is initially large and decreases toward convergence. A prominent merit of this algorithm is that it does not require any prior knowledge of the number and the shape of clusters, approximating the true distribution of the data. When used for clustering the data represented by a single variable, mean shift algorithm results in segmentation of the variable axis into several intervals (where the modes are centralized).

Suppose that mean shift algorithm segments the pth variable axis into Q intervals (for different variables, Q might be diverse), and n_q ($q = 1, \ldots, Q$) samples from different classes converged at the same mode that located in the qth interval, besides $n_q(k)$ samples from class k, then ISP score of gene p for class k ($k = 1, \ldots, K$), $ISP(k)$, is defined as follows:

$$ISP(k) = \sum_{q=1}^{Q} w_q \frac{n_q(k)}{n_q} = \sum_{q=1}^{Q} \frac{n_q(k)}{n(k)} \frac{n_q(k)}{n_q} \quad (k = 1, \ldots, K) \tag{4}$$

where w_q denotes the weight of purity of the qth interval to the total $ISP(k)$, which is determined by the percent of sample number belonging to class k in the qth interval to total sample number of class k, i.e., $w_q = (n_q(k)/n(k))$, where $n(k)$ symbolizes the number of samples in class k, i.e., $\sum_{q=1}^{Q} n_q(k) = n(k)$ and $\sum_{k=1}^{K} n(k) = N$. Based on the definition, large values of $ISP(k)$ imply that expression levels of gene p comprise intervals in which most of samples come purely from class k, indicating that this gene is useful for discriminating class k from the others. Thus, the genes with the largest $ISP(k)$ can be identified as the significant genes for class k. In general cases, several significant genes are selected for each class, and the optimal number, say J, of significant genes can be determined by trials and errors procedure or cross validation with varying number of significant genes. The simplest way is to choose the least number of significant genes when a model reaches desirable classification accuracy for the training set.

In addition, mean shift algorithm allows direct determination of the cluster centers for each class using the mode thus located. It offers a benefit for the following unimodal transform of the L selected genes that $L = J \times K$.

2.2. Unimodal transform of variables selected by ISP-based technique (UTISP)

Since the genes identified by ISP-based technique potentially exhibit multi-modal expressed patterns, straight generating decision rules using these genes may increase the overfitting or underfitting risk of the classification tree models because of the use of binary recursive partitioning strategies. Extracting their unimodal features for CART modeling can decrease these risks. Via mapping the significant genes into their higher dimensional spaces from the original one-dimensional ones using Gaussian functions,

the unimodal expression patterns of these genes can be disclosed and the unimodal featured variables can be extracted. This unimodal transform procedure in fact offers a flexible variable treating approach for CART to model microarray data, thus furnishes an approach toward the improvement of the performance of CART in generating classification rules. The output of unimodal transform of the lth variable can be represented as follows:

$$\mathbf{o}_l = \max\left\{\exp\left(-\frac{||\mathbf{x}_l - c_{l,1}||^2}{2{\sigma_{l,1}}^2}\right), \ldots, \exp\left(-\frac{||\mathbf{x}_l - c_{l,D_l}||^2}{2{\sigma_{l,D_l}}^2}\right)\right\},$$
$$d_\iota = 1, \ldots, D_\iota, \quad \iota = 1, \ldots, L \tag{5}$$

where c_{l,d_l} and σ_{l,d_l} denote to be the d_lth transform center and kernel width of gene l. Suppose gene l is one of the significant genes related to class k, then the mode centers where the entries of $\mathbf{x}_n(k)$ ($n = 1, \ldots, n(k)$), all training samples in class k, converged to in mean shift clustering, are to be the transform centers $\mathbf{c}_l = (c_{l,1}, \ldots, c_{l,D_l})$. Also, the number of centers D_l is automatically determined. It might vary for different genes. The widths $\boldsymbol{\sigma}_l = (\sigma_{l,1}, \ldots, \sigma_{l,D_l})$ in the present study is calculated by the distance of the nearest cluster center c_{l,d_l} to them [11,18]. To decrease the noises of data, a modified $\mathbf{c}_l$ is used in this study. The validity of c_{d_l} is measured by the percent of sample number belonging to class k to total sample number of class in the d_lth interval. Only if $(n_{d_l}(k))/n_{d_l} \geq per$, c_{d_l} is available. The parameter *per* also can be determined by sensitivity analysis.

Actually, for a significant gene related to class k that identified by ISP, the transform occurs at the mode centers of the intervals where its related class samples are centralized. Note that, the related samples have smaller distance from one of the mode centers than other samples. Then, in the feature space, samples from class k would reach biggish values on at least one new axis than other ones. Via assigning a sample the largest value it reached on any axis, one can obtain a feature variable with unimodal expression pattern, presenting larger values in all class samples of class k than in other ones. The feature variables essentially have enhanced ability in discriminating a certain class on one-dimensional space, which is much flexible for CART to generating decision rules.

2.3. UTISP for CART modeling (UTISP-based CART)

Classification trees are constructed on the output, $\mathbf{O} = (\mathbf{o}_1, \ldots, \mathbf{o}_L)$, of UTISP analysis using a CART algorithm [5]. The classic CART algorithm was popularized by Breiman et al. [5] CART is a nonlinear, parameter-free tree-building technique, constructing classification or regression trees for predicting continuous dependent variables (regression) or categorical predictor variables (classification). In the present study, CART is employed to generate classification decision rules for predicting the class of each sample from its known values of L explanatory variables $\mathbf{o}_l$ ($l = 1, \ldots, L$). Traditionally, there are two main steps in building a classification tree using a CART algorithm, growing a tree and pruning the tree.

CART grows trees using greedy binary recursive partitioning (splitting) of all the objects into smaller subgroups in a top-down manner, starting from the root node containing the entire training compounds until each node reaches completely homogeneity or a user-specified minimal sample number (i.e., node size) and becomes a terminal or leaf node. The binary recursive partitioning process, implemented by exhaustively searching all the unimodal transformed variables $\mathbf{o}_l$ ($l = 1, \ldots, L$) and all their possible values, respectively, formulates a series simple if/then rules according to the goodness-of-split criterion [5] Gini index or Twoing rule. And the classes are assigned to the nodes according to a rule that minimizes misclassification costs. A large tree $T_{\max}$ grown based on these rules consists of a number of intermediate, splitting nodes and a series of terminal nodes that represent homogeneous groups of observations in terms of the response variable. The explanatory variables appear in the consecutive splitting nodes in a hierarchy of decreasing explanatory power. In many cases, this tree is exposed to the risk of overfitting because of the exhaustive search during the growing process. Also, it is suffering from incomprehensibility. Pruning $T_{\max}$ back to an optimal sized tree to reduce these risks is inevitable. The most popular method for tree-pruning is the minimal cost-complexity pruning (MCCP), expressed as follows:

$$R_\alpha(T) = R(T) + \alpha\left|\tilde{T}\right| \tag{6}$$

where $R_\alpha(T)$ is a linear combination of the tree cost and complexity, representing the cost-complexity measure; $R(T)$ is the within-node sum of resubstitution error; $\alpha \geq 0$ refers to a complexity parameter and $\left|\tilde{T}\right|$ measures the tree complexity (i.e., the number of terminal nodes in the tree T). The more complex is the tree, the lower is $R(T)$, but at the same time, the higher is the penalty $\alpha\left|\tilde{T}\right|$ and *vice verse*. As proven by Breiman, α is a metric of additional accuracy introduced by a certain new bifurcation. The tree with α equaling to 0 refers to the unpruned largest tree $T_{\max}$ and α tends to increase from the leaf nodes to the root node. The MCCP tries to cut off the weakest branches of the tree in a bottom-up manner, gradually increasing α value. For a given α, among all subtrees of the same complexity, the optimal subtree $T \leq T_{\max}$ minimizing $R_\alpha(T)$ uniquely exists, as proven by Breiman. In such way, a sequence of nested subtrees is gained, from which a final approximately fit tree can be selected in terms of its best prediction accuracy either gained by cross validation method or pruning set technique.

A classification tree built based on the output of UTISP may display excellent performance in avoiding the risk of overfitting contrast to a tree straight built on the microarray data. Though the tree-pruning technique can decrease the overfitting risk, in classification analysis using microarray based gene expression signatures, the obtained model is hard to avoid overfitting because of lots of noises and redundant information rooting in exhaustively searching a great amount of irrelevant genes in microarray data. Identifying significant genes using ISP-based method could remove most irrelevant genes, providing the most discriminative variables by revealing the intrinsic relationships between with-class samples and between-class samples. In addition, after the unimodal transform the discriminative power of each gene in one-dimensional space would be enhanced, improving the performance of CART in construct a classification tree with desirable precision and generalization ability.

All the algorithms used in this study were written in Matlab 6.1 and run on a personal computer (Intel Pentium processor 4/2.80 GHz 1 GB RAM).

3. Results and discussion

3.1. Leukemia data

The performance of the proposed UTISP-based CART is first examined by leukemia data [14]. This data set is revisited here as a three-class classification problem, consisting of 72 samples including myeloid leukemia (AML), B-cell acute lymphoblastic leukemia (ALL) and T-cell ALLs. There are 38 samples in the training set and 34 samples in the test set. All samples are represented by 7129 human genes. Additional preprocessing steps were taken into this data set before standardization: (1) thresholding (floor of 100 and ceiling of 16,000), (2) filtering (exclusion of genes with max/min ≤ 5 and max $-$ min ≤ 500 across the samples), (3) base 10 logarithmic transformation [19]. This filtering resulted in 3571 genes. Then, the 3571 genes were scaled into [0.1, 0.9].

Table 1
Classification results of leukemia data using a UTISP-based CART compared with those obtained by *k*-NN, conventional CART, and BSS/WSS-based CART.

Method	No. of genes	Prediction accuracy	
		Training set	Test set
k-NN (*k* = 4)	3571		0.8529
CART	3571	0.9737	0.8529
BSS/WSS-based CART	160	0.9737	0.8235
UTISP-based CART	21	0.9737	0.9706

As a comparison, *k*-nearest neighbors algorithm (*k*-NN) was invoked to classify the three types of leukemias using the 3571 variables. According to the result of cross validation, *k* was set to be 4. Five test errors for ALL/AML problem were obtained by *k*-NN. The prediction accuracy for the 34 test samples is 85.29%. Then, CART was employed to construct the classification model. Considering the precision of the classification model and the complexity of the model, the splitting of a node would be stopped if there are less than 5 samples in this impure node. The prediction errors provided by the best pruned classification tree determined by cross validation (CV) are one sample for the training set and five samples for the test set, as shown in Table 1. The classification model built by a CART algorithm emerges series overfitting. The prediction accuracy for the training set is 97.37% (1 error among 38 samples), but the prediction accuracy for the test set is 85.29% (5 errors among 34 samples). Contrast with *k*-NN, CART did not improve the classification accuracy. A good many irrelevant, insignificant or redundant genes among the 3571 variables increased the risk of overfitting of the classification model. Extracting the most discriminative variables to build the CART model is a feasible approach to improve the performance of a model. The gene selection method based on ranking the ratio of between classes sum of squares to within class sum of squares for each gene (BSS/WSS) [19] was employed to pick out the relevant variables in the leukemia data set. BSS/WSS based method has been employed by many studies. As shown in Table 1, the best prediction results obtained by the classification tree models built based on different number of variables for the training set and the test set are 97.37% (1 error among 38 samples) and 82.35% (6 errors among 34 samples), respectively, when 160 genes were used to training the model. A high precise model was obtained, but the generalization of the model was very poor. Although in the method BSS/WSS-based CART, parts of redundant genes were eliminated, the generalization of the model built by BSS/WSS-based CART was not improved, indicating that the ability of BSS/WSS for identifying signification genes is limited and restricted the model in conquering the overfitting problem.

To improve the classification model, UTISP-based CART algorithm was applied to the leukemia data set. The parameters *h* (analysis bandwidth) and *per* (critical value determining the validity of transform center c_{d_l}) were determined by sensitivity analysis [20,21]. The sensitivity analysis of the parameter analysis bandwidth *h* is revealed in Fig. 1. It can be observed that if *per* takes 0.1 and *J* (the number of significant genes selected for each class) takes 7, the classification model reaches the desirable precision when $h \leq 0.25$. And the sensitivity analysis of the parameter *per* shown in Fig. 2 exhibits that when $per \geq 10\%$, the classification model reaches the best precision if *h* takes 0.25 and *J* takes 7. For ensuring the precision and the generalization ability of the models, *h* takes 0.25 and *per* takes 10% in the classification tree model of leukemia. The classification results obtained by UTISP-based CART are shown in Table 1. Taking advantage of unimodal transform of the variable selection based on ISP, the UTISP-based CART showed a good performance both for modeling and prediction. When *J* takes 7, the classification accuracy for the training sets is 97.37%, one error among the 38 training samples and the classification accuracy for the test set is 97.06%, one error among the 34 test samples, indicating that the UTISP-based CART yielded a classification tree with desirable precision and generalization ability. Via picking out a small number of the most discriminative variables of leukemias and unimodal transform of them, UTISP-based CART is efficient in conquering the overfitting problem which is very serious in conventional CART. Compared with CART and BSS/WSS-based CART as showing in Fig. 3, UTISP-based CART provides the same good performance for the training set but better performance for the test set, exhibiting that the proposed method has good precision in modeling and superior generalization in prediction. Significant genes of leukemias were well addressed by an ISP-based variable selection method, and unimodal transform of these genes further simplified the selected data, contributing to the superb performance of a classification tree model.

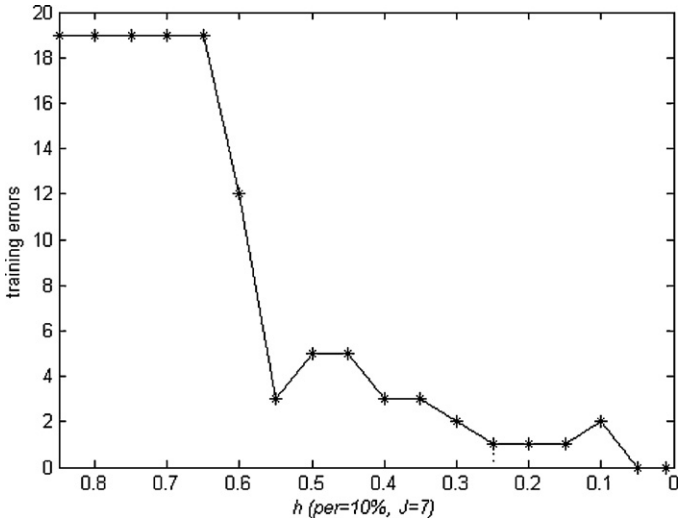

Fig. 1. The sensitivity analysis of analysis bandwidth *h* for leukemia data when using a UTISP-based CART method with *per* taking 10% and *J* taking 7.

3.2. Small, round blue-cell tumors data

For further testing the performance of UTISP-based CART, this novel strategy is applied to classify different kinds of small, round blue-cell tumors (SRBCTs) of childhood [15]. This data set with

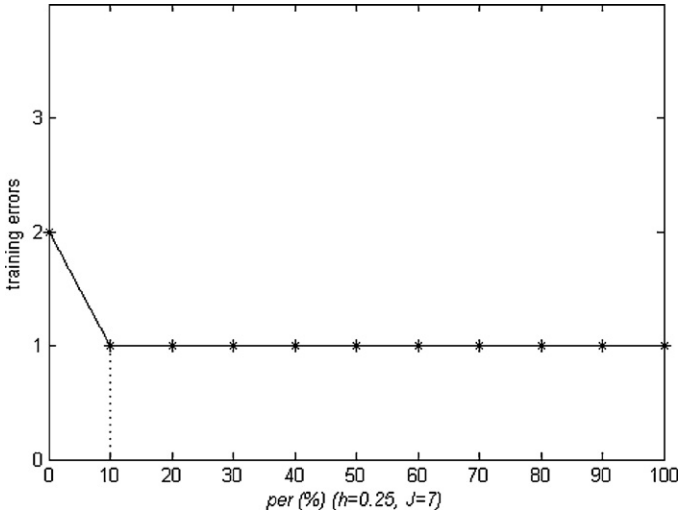

Fig. 2. The sensitivity analysis of *per* for leukemia data when using a UTISP-based CART method with analysis bandwidth *h* taking 0.25 and *J* taking 7.

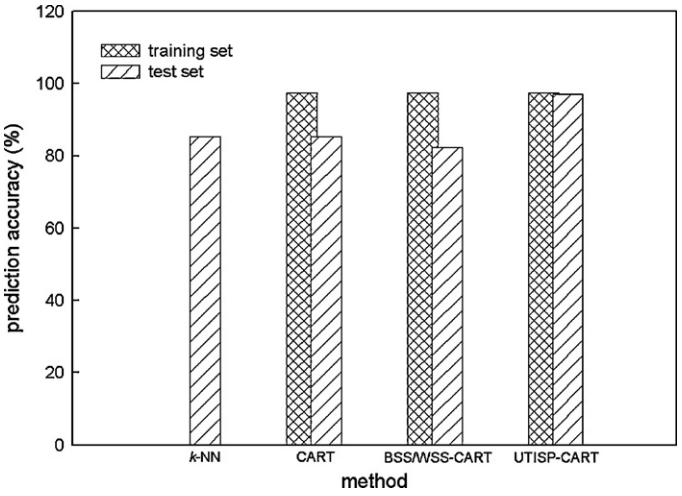

Fig. 3. The prediction accuracies of leukemia data obtained using *k*-NN, conventional CART, BSS/WSS-based CART, and UTISP-based CART.

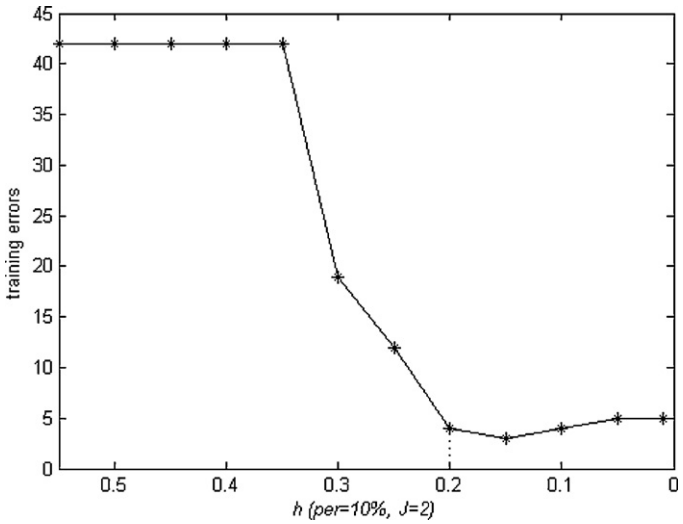

Fig. 4. The sensitivity analysis of analysis bandwidth *h* for SRBCTs data when using a UTISP-based CART method with *per* taking 10% and *J* taking 2.

83 samples includes four subsets of SRBCTs, neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS). Discrimination of these four types has presented a challenge due to their similar appearances in routine histology. In 2001, Khan and his co-workers successfully monitored the gene expression profiles of 6567 genes for these four types of malignancies using cDNA microarrays and reduced the number of genes to 2308 by quality filtering for a minimal level of expression [15]. In the present study, the 2308 genes are scaled into [0.1, 0.9].

To avoid a fortuitous choice of the training and the test sets, the original data set was randomly divided into five parts of roughly equal size. In each experiment, one part of them would be used as a test set and the rest making up a training set. The five different random combinations of training and test sets were investigated. As comparisons, *k*-NN, CART and BSS/WSS-based CART were also employed. The procedures were all repeated for the five combinations. Table 2 summarizes the statistical results of the mean classification accuracy of the five computations. The best result given by *k*-NN is a mean prediction accuracy of 77.80% for the test set when $k = 3$ that determined by cross validation results. The poor result reveals that *k*-NN failed to exhibit the relationship between different types of small, round blue-cell tumors. The mean classification accuracy of the five computations obtained by CART is 94.22% for the training sets and 79.54% for the test sets. High model precision was achieved. Nevertheless, the mean prediction accuracy for the test sets shows the models built by the conventional CART enduring serious overfitting. Contrast with conventional CART, the BSS/WSS-based CART improved the classification tree model slightly. The best result obtained by BSS/WSS-based CART has mean prediction accuracy 97.59% for the training sets and 84.43% for the test sets when the top 100 genes were used. Though BSS/WSS-based CART provided a better result than CART, the prediction accuracy of 84.43% for the test sets is undesirable yet. Due to limited variable selection ability of the variable method based on BSS/WSS, the overfitting problem of the classification trees is still serious, exposed by the low prediction accuracy for the test sets.

To improve the classification model, the UTISP-based CART algorithm was applied to predict the class type of small, round blue-cell tumors. According to the sensitivity analysis of the parameter analysis bandwidth *h* demonstrated in Fig. 4 and *per* in Fig. 5 respectively, it is clear that when *h* is smaller than or equal to 0.2, and *per* is greater than or equal to 10%, large changes in *h* or *per* result in a relatively small changes in the outcomes, so *h* takes 0.2 and *per* takes 10% in these models when *J* is 2. The mean classification results of the five computations as shown in Table 2 are 94.59% for the training sets and 91.59% for the test sets, respectively. In Fig. 6, it can be observed that compared with *k*-NN, CART and BSS/WSS-based CART, the newly proposed method presents the best performance: UTISP-based CART not only guaranteed the similar high precision for the models as CART and BSS/WSS-based CART, but also improved the generalization ability of the classification tree models. The prediction accuracy of 79.54% for the test sets obtained by CART and 84.43% obtained by BSS/WSS-based CART were improved to 91.59% by UTISP-based CART. It can be concluded UTISP-based

Table 2
Classification results of SRBCTs data using a UTISP-based CART compared with those obtained by *k*-NN, conventional CART, and BSS/WSS-based CART.

Method	No. of genes	Mean prediction accuracy	
		Training sets	Test sets
k-NN ($k = 3$)	2308		0.7780
CART	2308	0.9422	0.7954
BSS/WSS-based CART	160	0.9759	0.8443
UTISP-based CART	8	0.9459	0.9159

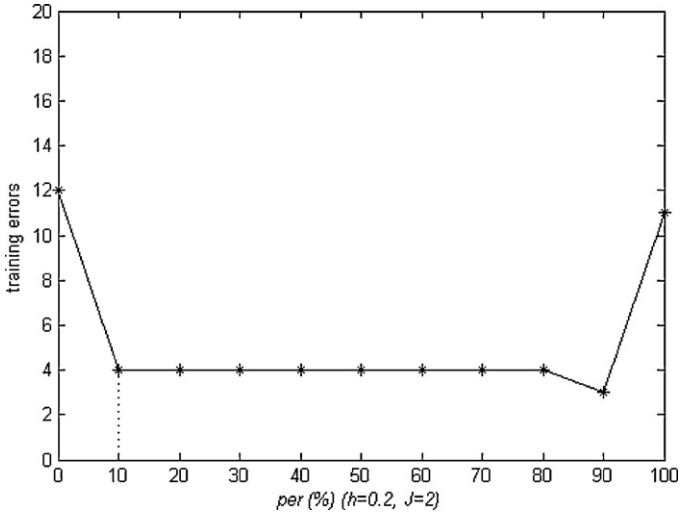

Fig. 5. The sensitivity analysis of *per* for SRBCTs data when using a UTISP-based CART method with analysis bandwidth *h* taking 0.2 and *J* taking 2.

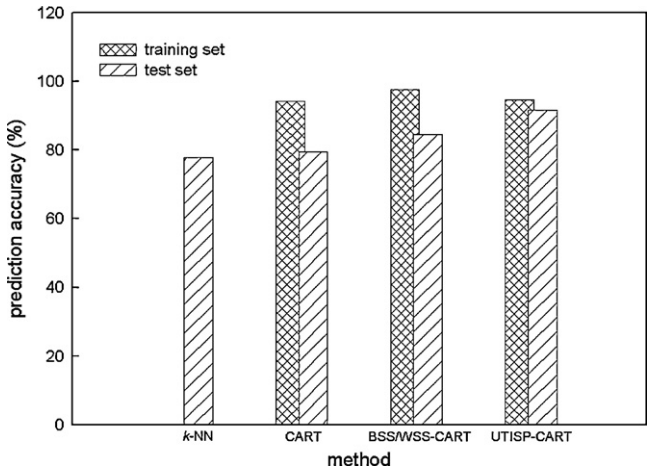

Fig. 6. The prediction accuracies of SRBCTs data obtained using *k*-NN, conventional CART, BSS/WSS-based CART, and UTISP-based CART.

CART improves the model prediction performance effectively. As a whole, the results revealed that the UTISP-based CART had the best performance rather than *k*-NN, CART and BSS/WSS-based CART on constructing the models for SRBCTs classification. The average results of five computations for the models built using UTISP-based CART expose that the good performance was not due to a fortuitous choice of the training sets and the test sets.

4. Conclusion

In this paper, unimodal transform of variables selected by interval segmentation purity for classification and regression tree modeling strategy has been developed for combating the challenge of high-dimensional microarray data with redundancy of variables. The development of ISP-based variable selection method provides an effective approach for significant gene identification. Unimodal transform of significant variables reduces the risk of overfitting of CART via extracting variable features. The performance of UTISP-based CART was evaluated by using two microarray data sets which reveals the proposed strategy is of great promise in generating classification trees with good stability and desirable performance in conquering overfitting for multi-class classification problems using microarray based gene expression signatures.

Acknowledgments

This work was supported by NSFC (21025521, 21035001 and 20875027), National Key Basic Research Program (2011CB911000), European Commission FP7-HEALTH-2010 Programme-GlycoHIT (260600), CSIRT Program and NSF of Hunan Province (10JJ7002).

References

[1] C.R. Andersson, M. Fryknäs, L. Rickardson, R. Larsson, A. Isaksson, M.G. Gustafsson, J. Chem. Inf. Model. 47 (2007) 239.
[2] E. Lee, H.Y. Chuang, J.W. Kim, T. Ideker, D. Lee, PLoS Comput. Biol. 4 (2008) 1.
[3] A.S. Leonardson, J. Zhu, Y.Q. Chen, K. Wang, J.R. Lamb, M. Reitman, V. Emilsson, E.E. Schadt, Hum. Mol. Genet. 19 (2010) 159.
[4] D. Baron, A. Bihouée, R. Teusan, E. Dubois, F. Savagner, M. Steenman, R. Houlgatte, G. Ramstein, Bioinformatics 27 (2011) 725.
[5] L. Breiman, J.H. Friedman, R.J. Olshen, C.J. Stone, Classification Regression Trees, Wadsworth, Pacific Grove, CA, 1984.
[6] J.S. Chang, R.F. Yeh, J.K. Wiencke, J.L. Wiemels, I. Smirnov, A.R. Pico, T. Tihan, J. Patoka, R. Miike, J.D. Sison, T. Rice, M.R. Wrensch, Cancer Epidemiol. Biomarkers Prev. 17 (2008) 1368.
[7] Y. Allory, C. Bazille, A. Vieillefond, V. Molinié, B. Cochand-Priollet, O. Cussenot, P. Callard, M. Sibony, Histopathology 52 (2008) 158.
[8] K. Srivastava, A. Srivastava, B. Mittal, Cancer 116 (2010) 3160.
[9] H. Sato, J.C. Grutters, P. Pantelidis, A.N. Mizzon, T. Ahmad, Am. J. Respir. Cell Mol. Biol. 27 (2002) 406.
[10] J. Katahira, H. Sugiyama, N. Inoue, Y. Horiguchi, M. Matsuda, N. Sugimoto, J. Biol. Chem. 272 (1997) 26652.
[11] L.J. Tang, W. Du, J.H. Jiang, H.L. Wu, G.L. Shen, R.Q. Yu, J. Chem. Inf. Model. 49 (2009) 2002.
[12] L.J. Tang, J.H. Jiang, H.L. Wu, G.L. Shen, R.Q. Yu, Talanta 79 (2009) 260.
[13] Y.Z. Cheng, IEEE Trans. Pattern Anal. Mach. Intell. 17 (1995) 790.
[14] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Science 286 (1999) 531.
[15] J. Khan, J.S. Wei, M. Ringnér, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, P.S. Meltzer, Nat. Med. 7 (2001) 673.
[16] K. Fukunaga, L.D. Hostetler, IEEE Trans. Inf. Theory 21 (1975) 32.
[17] Y. Cheng, K.S. Fu, IEEE Trans. Pattern Anal. Mach. Intell 7 (1985) 592.
[18] N. Benoudjit, C. Archambeau, A. Lendasse, J. Lee, M. Verleysen, Width optimization of the Gaussian kernels in radial basis function networks, in: ESANN'2002 Proceedings: European Symposium on Artificial Neural Networks, Bruges, Belgium, 24–26 April 2002, 2002, pp. 425–432.
[19] S. Dudoit, J. Fridlyand, T.P. Speed, J. Am. Stat. Assoc. 97 (2002) 77.
[20] W.Q. Lin, J.H. Jiang, Y.P. Zhou, H.L. Wu, G.L. Shen, R.Q. Yu, J. Comput. Chem. 28 (2007) 519.
[21] L.J. Tang, Y.P. Zhou, J.H. Jiang, H.Y. Zou, H.L. Wu, G.L. Shen, R.Q. Yu, J. Chem. Inf. Model. 47 (2007) 1438.