Artificial Intelligence in Medicine 48 (2010) 129–137

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Mixture classification model based on clinical markers for breast cancer prognosis

Tao Zeng, Juan Liu *

School of Computer, Wuhan University, Wuhan 430079, China

ARTICLE INFO

Article history:
Received 22 August 2008
Received in revised form 9 July 2009
Accepted 20 July 2009

Keywords:
Mixture classification model
Breast cancer prognosis
Rough set
Support vector machine
Clinical marker

ABSTRACT

Objective: Accurate cancer prognosis prediction is critical to cancer treatment. There have been many prognosis models based on clinical markers, but few of them are satisfied in clinical applications. And with the development of microarray technologies, cancer researchers have discovered many genes as new markers from the gene expression data and have further developed powerful prognosis models based on these so-called genetic biomarkers. However, the application of such biomarkers still suffers from some problems. The first one is there are a great number of genes and a few samples in the gene expression data so that it is difficult to select a unified gene set to establish a stable classifier for prognosis. The second one is that, due to the experimental and technical reasons, there are existing noises and redundancies in gene expression data, which may lead to building a prognosis predictor with poor performance. The last but not the least one is the microarray experiments are so expensive currently that it is hard to obtain abundant samples. Therefore, it is practical to develop prognosis methods mainly based on conventional clinical markers in real cancer treatment applications. This paper aims to establish an accurate classification model for cancer prognosis, in order to make full use of the invaluable information in clinical data, especially which is usually ignored by most of the existing methods when they aim for high prediction accuracies.

Methods: First, this paper gives the formal description of general classification problem, and presents a novel mixture classification model to make full use of the invaluable information in clinical data, which is similar to the traditional ensemble classification models except for putting strict constraints on the construction of mapping functions to avoid voting process. Then, a two-layer instance of the proposed model, named as MRS (Mixture of Rough set and Support vector machine), is constructed by integrating rough set and support vector machine (SVM) classification methods, in which, the rough set classifier acts as the first layer to identify some singular samples in data, and the SVM classifier acts as the second layer to classify the remaining samples. Finally, MRS is used to make prognosis prediction on two open breast cancer datasets. One dataset, denoted as BRC-1 hereafter, is a high quality, publicly available dataset of 97 breast cancer tumors of node-negative patients. The other, denoted as BRC-2 hereafter, uses baseline human primary breast tumor data from LBL breast cancer cell collection containing 174 samples.

Results: We have done two experiments on BRC-1 and BRC-2, respectively. In the first experiment, the BRC-1 dataset is divided into train set with 78 patients (34 ones belonging to poor prognosis group and 44 ones belonging to good prognosis group) and test set with 19 patients (12 ones belonging to poor prognosis group and 7 ones belonging to good prognosis). After trained on the train set, the MRS can correctly classify all the 12 patients with poor prognosis, and 6 of 7 patients with good prognosis in the test set. The results are better than previous researches, even better than the 70-gene based biomarkers. And in the second experiment, we construct the classifiers using BRC-2 dataset, and compare MRS with other representative methods in Weka software by 5-fold cross-validation, and comparison results show that MRS has higher prediction accuracy than those methods.

Conclusions: The proposed mixture classification model can easily integrate methods with different characteristics. It can overcome the shortcomings of traditional voting-based ensemble models and thus can make full use of the information in clinical data. The experimental results illustrate that our implemented MRS classifier can predict the breast cancer prognosis more accurately than previous prognostic methods.

© 2009 Elsevier B.V. All rights reserved.

* Corresponding author. Tel.: +86 27 6277 3741; fax: +86 27 6877 8582.
E-mail addresses: zt_2003@163.com (T. Zeng), liujuan@whu.edu.cn (J. Liu).

0933-3657/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2009.07.008

1. Introduction

In clinical applications, the physicians usually need to make feasible judgments on the therapies to patients by predicting therapy response, prognosis, metastatic phenotype, etc. before they carrying out practical cancer treatments [1]. Particularly for breast cancer patients with lymph node negative, it would be helpful to predict the presence of distant metastasis, which is noted as poor prognosis based on the primary tumor. Therefore, strong prognostic markers are needed to support the physicians in making treatment decisions by dividing breast cancer patients into different risk groups [2].

There are already many conventional prognostic markers used in clinical applications, such as clinical (age), histopathological (lymph node status, tumor size, histological grade), cell biological oestrogen receptor (ER) and progesterone receptor (PgR). Since some of the cancer researchers consider them not to be sufficient for predicting cancer prognosis [1], they devote to discovering more powerful prognostic markers from developing gene expression data. In fact, during the past decade, microarray technologies have had a great impact on cancer researches such as disease class discovery and the prediction of diagnosis, prognosis or treatment response [1–10]. Especially, genetic biomarkers extracted from gene expression data have shown great potential in breast cancer prognostic prediction [1,6,11]. The gene expression data indeed have more genetic information than clinical data, however, current classification methodologies still have application bottleneck of using these information in clinical outcome prediction, just due to high dimension of data, much more noises to signal, no supporting rules for judgment, and so on [11]. Moreover, about half of all new breast cancers are diagnosed in the third world, where the analysis of prognostic factors needs to be inexpensive, be easy to perform, and have a good reproducibility [12]. While the microarray experiments are expensive nowadays, it is impractical to use genes as the prognostic markers of breast cancer at most times. In one word, it is emergent to develop powerful methods to make accurate cancer prognosis prediction based on clinical data.

As known to all, recent researches on clinical outcome predictions suggest the ensemble methodology in machine learning community would be a possible solution for promoting the efficiency and robustness of the cancer prognostic system in silicon. For example, in order to increase the accuracy of decision support system for early detection and diagnosis of breast cancer, West et al. investigated different strategies for selecting classification models from a pool to form the ensemble classifiers [13]; Geurts et al. adopted majority voting strategy and accuracy weighted voting scheme to automatically determine the proteomic biomarkers and predictive models based on four parallel methods (bagging, random forests, extra-trees and boosting) [14]; and in order to enhance the robustness of microarray data classification, Peng first trained a pool of candidate base classifiers from separate sample subsets, then selected appropriate base classifiers, and finally constructed the classification committee from the clustered classifiers [15]; Ge and Wong built the decision tree based ensemble classifier with unweighted voting strategy [16]; Koziol et al. investigated that the ensemble tree classifiers could robustly and reproducibly identify different subgroups of patients [17].

However, the traditional ensemble methods still have their limits in clinical application. First, the voting scheme, either majority voting or weighted voting, is a kind of strategies with bias, such that the ensemble classification model is usually inclined to accept clinical outcome predictions with more voting supports, which may ignore the special samples in clinical prognosis. Second, the construction methods of candidate classifiers or the types of classifiers in ensemble pool are usually similar. Since every classification method has its own weakness, the ensembles from the similar types of classifiers would also inherit such drawbacks. Therefore, this paper designs a novel *mixture classification model* for breast cancer prognosis application. The term "mixture", different from normally used "ensemble" or "integrate", emphasizes that such method will avoid unnecessary voting process in ensembles construction and can combine intrinsic different classification methods such as rules-based methods and non-rules-based technologies.

The remains of this paper are organized as follows: Section 2 first introduces the formal description of the general classification problem, and then proposes the mixture classification model; Section 3 presents the construction of the mixture classifier based on rough set and support vector machine (SVM); the experiments and results of the constructed mixture classifier on breast cancer datasets are detailed in Section 4; Section 5 discusses the significant clinical meaning revealed from the rules composed of the first layer of relative mixture classifier; and the conclusions are drawn in the last section.

2. Mixture classification modeling

2.1. Description of classification problem

First of all, we formally describe the general classification problem. Let the object (e.g. sample, patient) set is $\Omega = \{\pi_i\}_{i=1}^{I}$ and the class (e.g. response, metastatic phenotype) set is $T = \{t_j\}_{j=1}^{J}$, each object in Ω corresponds with a class type in T. This object–class relationship can be modeled by function mapping in mathematical term, and the so-called *classification function* Φ can be described as:

$$\Phi(\pi_i) = t_j, \quad \pi_i \in \Omega, \quad t_j \in T$$

Obviously, some different objects possibly have the same class type, or in other words, these objects belong to the same class. This concept can be formalized, as:

$$\Phi(\Omega_j) = t_j, \quad \Omega_j \subseteq \Omega, \quad t_j \in T$$

where Ω_j, called as class field, is the subset of Ω, and satisfies:

$$\begin{cases} \forall \pi_i \in \Omega_j, \quad \Phi(\pi_i) \equiv t_j \\ \forall \pi_i \in \Omega - \Omega_j, \quad \Phi(\pi_i) \neq t_j \end{cases}$$

It is required that all the class fields have the properties of disjointness and completeness, which means:

$$\begin{cases} \Omega_m \cap \Omega_n = \phi, \quad \forall m, n \in [1, J], \quad m \neq n \\ \bigcup_{k=1}^{J} \Omega_k = \Omega \end{cases}$$

The property of disjointness guarantees that any object has only one class type, and the property of completeness makes sure that every object has its own class type. According to above definitions, each known object has an exclusive and formal class type by default. However, there may be some unknown objects whose class types are not very clear. Either their class types are in the class set T, or they even belong to some undiscovered class. Then T can be extended to $T^* = T \cup \{\mu\}$, where μ stands for a new class type like "unknown". So, we always have a class type mapping function as:

$$\Phi(\Omega, T) = \Phi(\Omega, T^*) = \begin{cases} \Phi(\Omega_j) = t_j, \quad \forall 1 \leq j \leq J \\ \Phi(\Omega_{J+1}) = \mu \end{cases}$$

For each object set Ω and its corresponding class set T, the above mapping function Φ always exists whether there are some new classes or not. We call Φ as a *native classification function*.

Assume the objects set is Ω and the class set is T, the objects with known class types compose the set $R(\subset \Omega)$, and the remains compose the set $P(= \Omega - R)$. R is also known as the *train set*, and P is

as the *prediction or test set*. The purpose of classification modeling is to find a mapping Φ^*, which has minimal difference with the native classification function $\Phi(\Omega,T)$. There are many kinds of measurements to evaluate the difference between such two functions. In case of our research, we redefine three widely accepted measurements as following.

Definition 1. The accuracy measurement is defined as:

$$A(\Phi,\Phi^*|\Omega)=\frac{|\{\pi_i|\Phi^*(\pi_i)=\Phi(\pi_i),\quad \pi_i\in\Omega\}|}{|\Omega|}$$

Definition 2. The partial accuracy measurement is defined as:

$$PA(\Phi,\Phi^*|\Omega,t_j)=\frac{|\{\pi_i|\Phi^*(\pi_i)=\Phi(\pi_i)=t_j,\quad \pi_i\in\Omega\}|}{|\{\pi_i|\Phi^*(\pi_i)=t_j,\quad \pi_i\in\Omega\}|}$$

In fact, the accuracy measurement is just the *ACC* indicator for all classification problems [18]. The partial accuracy measurement is equivalent to the *sensitivity* or *specificity* evaluations when it is used in two-class problems [19]. And another usually applied Matthew's correlation coefficient (*MCC*) [20] can also be redefined as following correlated accuracy measurement when J equals to two.

Definition 3. The correlated accuracy measurement is defined as:

$$C(\Phi,\Phi^*|\Omega)=\frac{\prod_{j=1}^{J}|\{\pi_i|\Phi^*(\pi_i)=\Phi(\pi_i)=t_j,\quad \pi_i\in\Omega\}|-\prod_{j=1}^{J}|\{\pi_i|\Phi^*(\pi_i)\neq\Phi(\pi_i)=t_j,\quad \pi_i\in\Omega\}|}{\sqrt{\prod_{j=1}^{J}|\{\pi_i|\Phi(\pi_i)=t_j,\quad \pi_i\in\Omega\}|\prod_{j=1}^{J}|\{\pi_i|\Phi^*(\pi_i)=t_j,\quad \pi_i\in\Omega\}|}}$$

In actual applications, the classifier is always constructed from the train set R, thus the whole objects set Ω will correspondently be replaced by the prediction set P in order to avoid self-examination on R when computing the above three measurements.

2.2. Mixture classification modeling

Up to now, there are many well-known classification methods/strategies, such as decision tree, random tree, rough set, fuzzy set, artificial neural networks, Bayesian network, support vector machine, and so on [21]. These methods can be divided into two groups with different classification principles. One kind of them is represented by the methods as decision tree, random tree, rough set, and fuzzy set. When used in two-class identification problems, these methods split the train/predict samples into three parts: positive samples, negative samples (both of them are also called as type-known samples), and type-unknown samples. There are two important indicators for evaluating these methods, *i.e.*, *coverage* and *accuracy*. The *coverage* is the quotient of the number of predicted type-known samples and the number of all samples; while the *accuracy* is the quotient between the number of correctly predicted samples and the number of predicted type-known samples. It should be noted that the optimal strategy (as pruning used in decision tree) used these methods for promoting prediction accuracy usually decreases the coverage, which will reduce the efficiency of classifiers due that many samples remain unknown and are useless for supporting decisions (Fig. 1(A)). The second kind of methods such as artificial neural networks, Bayesian network, and support vector machine, always tries to assign a class type to each sample although it may be untrue (Fig. 1(A)). These methods normally find out a few existing singular samples that might cause the misclassification, and kick them out ahead in order to obtain higher prediction accuracies. However, in most clinical applications the data is invaluable as the patients/samples are few or it is obtained via huge expense. Remaining some samples out of consideration is so wasteful that much information in the abnegated data is lost during classification modeling.

A possible solution to above problems is the strategy of ensemble classification [11,13–17] or meta-analysis [22,23] (Fig. 1(B)), which can be briefly described as below. Given the object set Ω, the class set T, we want to search a series of mapping functions $\{\Phi_k^*(\Omega,T)\}_{k=1}^{K}$ and their ensemble function $\Phi^\infty(\Phi_1^*,\Phi_2^*,\ldots,\Phi_K^*)$, so that the difference between Φ^∞ and the native classification function $\Phi(\Omega,T)$ is minimum. The measurements of such difference can be the same as above three definitions. In fact, the series of mapping functions can be usually obtained through some well-known classification methods listed above, while the main hurdle in ensemble classification modeling is the construction of ensemble function. The widely used ensemble strategy is based on voting. However, this strategy tends to be of bias for the truth may only be supported by a small number of votes. In the case of cancer prognosis, where patients should be considered individually, the ensemble strategy based on voting is obviously not appropriate.

Inspired by one indistinguishable case in a 2-dimension space (Fig. 2), this paper proposes a novel ensemble-like method, which puts a strict constraint on the construction of mapping functions to avoid voting process. In Fig. 2(A), any single classifier (represented by dot line) can only achieve 80% accuracy at most, for there is no line to divide the complete sample space (object set) into two parts such that each part covers one class field. For example, if the left area of sample space is considered to be the field of Class 1, and the right one is the field of Class 2, two samples would always be misclassified. Although someone could find another line (new classifier) to recognize such two samples, this new classifier will still misclassify some other samples. So in previous researches, these two samples are either kept to be misclassified, or be directly omitted from the data before classification. However, just mentioned above, these strategies are not acceptable in many applications such as cancer prognosis, where samples are so precious that much information may be contained in the abnegated data (those misclassified or omitted samples).

Our basic idea is shown in Fig. 2(B). First, we build a classifier to classify samples in some sample subspace (illustrated as the dotted ring), leaving all samples out of the subspace unclassified. Then we construct another classifier to classify the remaining samples. By this way, all samples in the whole space are expected to be correctly classified.

Formally, suppose the native mapping function can be divided into several parts such that

$$\Phi(\Omega,T)=\sum_{k=1}^{K}\Phi_k(\Omega_k)$$

and

$$\Phi_k(\Omega_k)=\begin{cases}\Phi(\pi_i)=t_j, & \pi_i\in\Omega_k\\ \mu, & \pi_i\in\Omega-\Omega_k\end{cases}$$

where the $\{\Omega_k\}_{k=1}^{K}$ also have the properties of disjointness and completeness, $\{\Phi_k(\Omega_k)\}_{k=1}^{K}$ are called as *component functions*. For every object, only one of K component functions is used to determine its class type, so all objects should be classified without voting. Correspondently, we can construct an ensemble-like or mixture classification model by building mapping functions which have the minimal differences with the component functions. Given

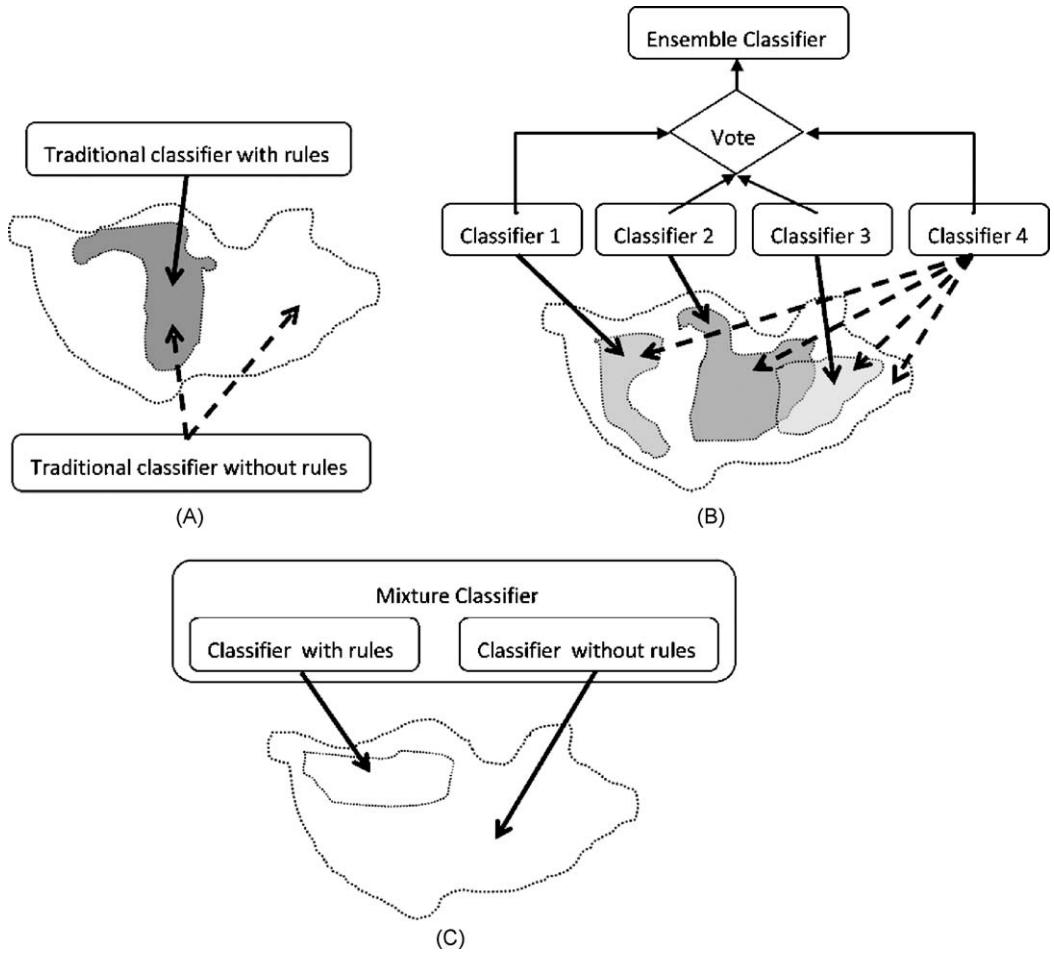

Fig. 1. The different construction strategies of classification models.

the object set Ω, the class set T, and the known class types of some objects in Ω, we first build a series of mapping functions $\{\Phi_k^*\}_{k=1}^K$, so that the difference between every pair of Φ_k^* and $\Phi_k(\Omega_k)$ is minimum according to above introduced measurements. Then the mixture classification function will have the form

$$\Phi^* = \sum_{k=1}^{K} \Phi_k^*$$

Obviously, in order to construct the mixture classification model, the native mapping function should be divided into at least two component functions. However, it is hard to guarantee the disjointness property when there are more than two component functions. So, in our implementation, we use only two components shown as in Fig. 1(C).

3. Construction of the mixture classifier

To implement our proposed mixture classification model described in Section 2, we construct the mixture classifier by combining rough set and SVM classifiers. The rough set classifier guarantees to identify some singular samples, it does not need to cover the whole train set but must have high prediction accuracy on those samples. This feature is called as "low coverage but high accuracy". While the SVM classifier is responsible to cover remain samples with enough accuracy.

3.1. Rough set reduction

Suppose every sample has n properties, the train set R can be represented as a decision table $D[p_1,p_2, \ldots, p_n]$, in which, each

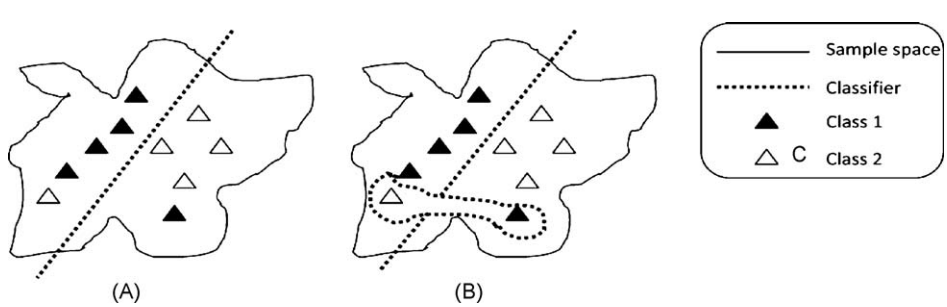

Fig. 2. An inspiring example of our mixture classification idea.

row represents a sample and each column represents a property. The element D_{ij} represents the value of sample i on property j. Usually, some of the properties are unhelpful or even harmful for correctly classifying the samples due to noise or redundancy. In order to winkle out the useless properties, it is necessary to do rough set reduction before building the classifier. So far, there are several methods to realize the rough set reduction on a decision table, such as dynamic reduction, exhaustive calculation, Johnson's algorithm, genetic algorithm, and so on [24]. In this paper, we use the dynamic reduction to get multiple dynamic reducts. A dynamic reduct with $m(<n)$ properties is denoted as $D[q_1,q_2,\ldots,q_m]$, which is a subtable of the original decision table. Every dynamic reduct has a support value denoting that how many samples obey the compatible rules on this decision subtable.

Ideally, the rough set classifier should recognize some singular samples, meanwhile, every extracted rules should cover as many samples as possible (one rule for one sample is too trivial in many applications). So, the threshold of support value V_{st} is usually needed to filter out the trivial dynamic reducts with too small support values. And then, the left dynamic reducts will be used as the candidates to construct rough set shrinking classifiers as described below.

3.2. Rough set shrinking classifier construction

Several candidate dynamic reducts are generated during the process of rough set reduction. For every dynamic reduct D_k, we first construct the related rough set classifier $R(D_k)$ using standard voting strategy and covering algorithm [24] (we also tried other methods such as exhaustive algorithm, genetic algorithm, and LEM2 algorithm, but the results are not better than the one used here in our experiments). According to the principle of "low coverage but high accuracy", we then filter the rules in $R(D_k)$. Instead adopting the time-consuming strategy of filtering the rules one by one, we perform groups filtering. If $R(D_k)$ does not satisfy the shrinking criterion, we extract $R_1(D_k)$ from $R(D_k)$ by deleting the rules with support values equal to one; and if $R_1(D_k)$ is still not acceptable, we further extract $R_2(D_k)$ from $R_1(D_k)$ by deleting the rules with support values equal to two. This iteration will continue until the first coming of $R_s(D_k)$ which satisfies the shrinking criterion or has an empty rule set. $R_s(D_k)$ is therefore regarded a candidate rough set shrinking classifier $R(D_k)$. Thus we can obtain multiple candidate classifiers, each for one dynamic reduct. Finally, we simply choose one with the largest coverage as the final rough set shrinking classifier.

In our implementation, rough set reduction and shrinking classifier construction are done by the Rough Set Exploration System 2.2.2 [24].

3.3. C-SVM classifier construction

In our work, the rough set shrinking classifier is expected to recognize some singular samples, and the other samples will be identified by C-SVM classifier. We use LibSvm package [25] to construct the classifier. The data is scaled by scale subprogram. The kernel function RBF and cross-validation method are used; and the parameters c and g are optimized by grid subprogram. All the properties of samples are under consideration without feature selection.

3.4. Mixture classifier construction

According to our experience, for a new sample, one usually first tries to find some special experience represented as rules helping to identify its class type. If the class type of this sample is not convinced, one then tries his non-rules-based experience to give a judgment. Thus the mixture classifier, noted as MRS (Mixture of Rough set and SVM), can be constructed as a two-layer model: the upper layer is rough set shrinking classifier and the lower layer is C-SVM classifier. For a new sample, the upper layer will first try to assign a class type with some supporting rules to it. If its class type is unknown, the second layer will then identify it by assign a class type to it. Fig. 3 shows the construction and prediction processes of MRS. Where the special rules of the mixture classification model (rough set shrinking classifier) may only validate for several singular samples. If the new sample is not a possible singular sample, the rough set shrinking classifier would assign it an unknown class type, so that it will to be classified in a non-rules-based system realized by C-SVM.

4. Experiments and results

In order to evaluate the effectiveness of our proposed method, we build the mixture classifiers based on the conventional clinical prognostic markers from breast cancer data, and compare them with other methods. We have done two kinds of experiments on two different datasets, respectively. The first experiment is done by using a high quality, publicly available dataset of 97 breast cancer tumors of node-negative patients, for which both clinical data and gene expression data have been offered, and we call it as BRC-1 dataset. The second experiment uses baseline human primary

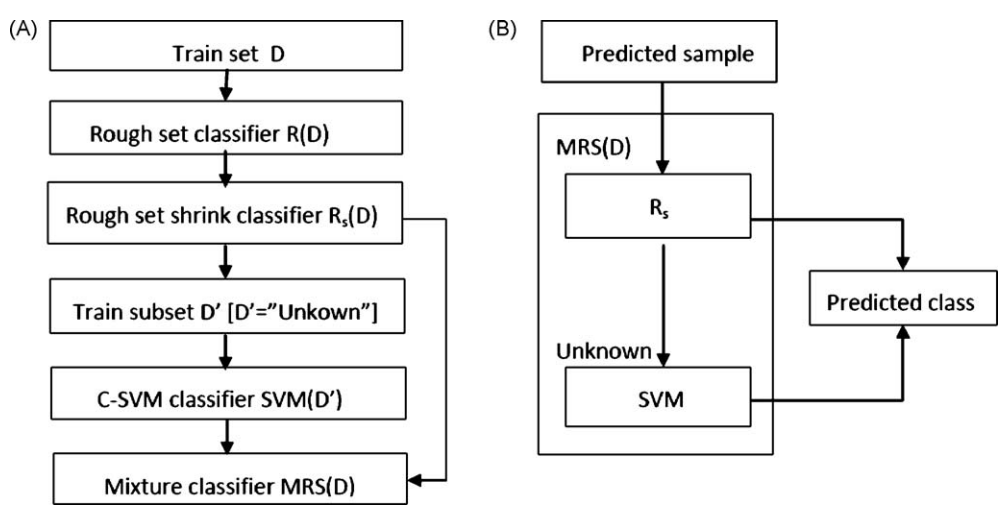

Fig. 3. The construction and prediction processes of MRS classifier.

breast tumor data from LBL breast cancer cell collection, which contains all the clinical information on the tumors reported in the paper of Chin et al. [26]. This dataset, named as BRC-2, contains 174 samples with 17 clinical parameters.

4.1. MRS on BRC-1 dataset with independent prediction test

Previous researches on BRC-1 include the traditional clinical indexes [12], the genetic biomarkers [1], and even the clinicogenomic markers based on clinical data and gene expression data [11]. BRC-1 is available from the Integrated Tumor Transcriptome Array and Clinical data Analysis database (ITTACA (2006)) [27]. It consists of train set and prediction set. The train set has 78 patients, in which, 34 patients who suffer disease recurrence in 5 years belong to the poor prognosis group, and other 44 patients belong to the good prognosis group without metastasis appearing at least 5 years. The prediction set contains 19 patients with 12 poor prognosis and 7 good prognosis patients, respectively. Every patient is described by seven clinical variables: age, diameter, tumor grade, oestrogen receptor (ERp) status, progesterone receptor (PRp) status, the presence of angioinvasion, and the presence of lymphocytic infiltration. Besides, among all patients, 68 are ER positive (+) (27 with metastasis and 41 without) and 29 are ER negative (−) (19 with metastasis and 10 without). Recent discoveries show that ER+ and ER− tumors represent two distinct disease phenotypes respectively [28], so it strongly suggests that the prognostic predictors would better be constructed separately for patients with ER positive or negative, which has already had an attempt study [12]. Here, we also follow the similar line to evaluate MRS in such prognosis applications.

In the dynamic reduction process, we first obtained 25 reducts, then we sorted them in descending order according to their support values. After filtering out those with support value fewer than five, we got six reducts listed in Table 1, which were to be used to build their related candidate rough set classifiers later. In Table 1, the first column shows the clinical parameters in each reduct, and the second column offers the support values of reducts. After the shrinking process, only the classifier, based on reduct {diameter (mm), ERp, PRp}, is successfully extracted as the most favorable rough set shrinking classifier. It can correctly recognize the class types of some singular patients (here the class type is with metastasis or not) and assign unknown class type to other patients. This classifier is represented as a rule set with rules listed in Table 2 (where each row represents one rule, with condition on the left, deduction on the right of the arrow), and indicates obvious clinical meaning of the classification.

The patients that are uncovered by the rule set and thus assigned the unknown class type, form a new train set for C-SVM construction. As RBF is used to be the kernel function, several c and g parameter pairs are used to produce candidate C-SVM classifiers. We chose the one with the best prediction performance.

In order to evaluate the performance of MRS, we compared it with other four methods: C-SVM, NPI, 70-gene biomarker and BPIM [1,11]. In particular, for the C-SVM method, we have built two classifiers respectively by using the total train set and by using the remaining train data after the construction of rough set shrinking classifiers. The results of comparison are listed in Table 3. Where, the first column shows different used classification methods; the second column is the area under ROC curve (AUC) for each classifier on BRC-1; the third column lists the number of patients predicted to be poor prognosis; and the last column shows the number of patients predicted to be good prognosis.

In previous studies [1,11], the constructed classifiers for original classification problem on whole 97 patients can also be directly used for classification problem on ER+ or ER− patients respectively without any modifications. That is because those methods based either on the genetic biomarkers or on the clinicogenomic markers does not make use of the information of ER at all. However, our constructed mixture classifier based on all patients needs ERp clinical parameter, whereas the construction classifiers based on ER+ or ER− patients must ignore ERp parameter for they should have pre-experiences on ERp before the modeling process.

Therefore, we constructed two classifiers on ER+ and ER− data sets separately. Taking ER+ classifier as an example, its train/predict set contains all the ER+ patients in original train/predict set, thus both the mixture classifier and the independent C-SVM classifier are constructed and tested on these new sets. The experiment results for two datasets are listed in Tables 4 and 5, respectively.

Table 1
The list of candidate reducts.

Reduct	Support
{age, diameter (mm)}	[60 subtable(s)]
{grade, age, ERp}	[20 subtable(s)]
{age, ERp}	[10 subtable(s)]
{diameter (mm), ERp, PRp}	[9 subtable(s)]
{grade, age, PRp}	[7 subtable(s)]
{grade, diameter (mm), ERp}	[6 subtable(s)]

Table 2
The rules representing favorite rough set shrinking classifier.

(diameter (mm) = 9) ≧ (metastasis = 1 [2])
(diameter (mm) = 10) ≧ (metastasis = −1 [4])
(diameter (mm) = 12) ≧ (metastasis = −1 [2])
(diameter (mm) = 27) ≧ (metastasis = 1 [2])
(diameter (mm) = 35) ≧ (metastasis = 1 [2])
(PRp = 10) ≧ (metastasis = −1 [4])
(PRp = 20) ≧ (metastasis = 1 [2])
(PRp = 30) ≧ (metastasis = −1 [2])
(PRp = 40) ≧ (metastasis = −1 [3])

Here metastasis have two values 1 and −1 used in experiments which represents recurrence and free, respectively. And the values in square brackets are support values for the rules.

Table 3
The comparison of different prognostic methods on BRC-1.

All97	AUC	Metastasis	Free
C-SVM[a]	0.74 (0.68)	9/12 (12/12)	6/7 (6/7)
NPI[b]	None	9/12	2/7
70-gene[b]	0.851	12/12	2/7
BPIM[b]	0.845	11/12	2/7
MRS[c]	0.70+	12/12	6/7

[a] C-SVM is built on two different train sets and tested on the same 19 patients prediction set. One train set is the original one with 78 patients, with result shown outside the bracket. The other train set contains patients with unknown class type assigned by rough set classifier, and its performance is recorded in bracket.
[b] These three methods have already been tested on this data set in research [11].
[c] For this method, its AUC only considers its component SVM classifier. So its actual AUC should be larger than the value shown here.

Table 4
The prognostic validation on ER+ sub-dataset.

ER+	AUC	Metastasis	Free
Others[a]	0.66–0.85	None	None
C-SVM	0.85 (0.72)	3/6 (3/6)	6/6 (6/6)
MRS	0.72+	3/6	6/6

[a] Others stand for the methods tested in study [12].

Table 5
The prognostic validation on ER− sub-dataset.

ER−	AUC	Metastasis	Free
Others[a]	0.19–0.61	None	None
C-SVM	0.68 (0.63)	6/6 (6/6)	0/1 (0/1)
MRS	0.63+	6/6	0/1

[a] Others stand for the methods tested in study [12].

Table 6
The comparison among MRS and other methods in Weka on BRC-2.

Methods	ACC	Specificity	Sensitivity	MCC
NaiveBayes[a]	73.58%	77.78%	60.00%	0.3464
SVM[a]	72.64%	75.00%	61.11%	0.2850
SVM[b]	78.34%	80.98%	78.34%	0.5120
J48[a]	70.75%	76.25%	53.85%	0.2797
RandomForest[a]	67.92%	71.91%	47.06%	0.1503
AttributeSelectedClassifier[a]	74.53%	76.14%	66.67%	0.3471
MRS	89.24%	87.21%	91.59%	0.7530

[a] These methods are offered in Weka software and trained in default parameters.
[b] This method is based on Libsvm with parameter optimization.

4.2. MRS on BRC-2 dataset with cross-validation

In the second experiment, after deleting the samples with missing values in BRC-2, 106 samples are used, where 33 samples with survival time larger than 5 years belong to good prognosis, and 73 samples with survival time fewer than 5 years belong to poor prognosis. As there is no independent prediction set offered, we evaluated the mixture classification model on this dataset by 5-fold cross-validation method just as previous studies, in order to compare it with other methods in Weka software [21]. The comparison results are shown in Table 6, where the evaluation indicators of ACC (accuracy), specificity, sensitivity, and MCC (Matthew's correlation coefficient) are displayed.

In order to investigate what clinical parameters might cause the difficulty in cancer prognosis, we observe the parameters involved in the rough set rules in every MRS. We run the 5-fold cross-validation 10 times, and obtained 50 classifiers. We counted the frequency for each clinical parameter in the classifiers, based on the assumption that the more frequently the parameter appears, the more significantly it is related to singular clinical outcome. The statistical results are shown in Fig. 4, where the x-axis represents 17 clinical parameters and y-axis indicates the total appearing times of each parameter.

5. Discussions

5.1. High accuracy and clinical meaningful rules of MRS on BRC-1 dataset

In the experiments on BRC-1 datasets, there are some notable results. The rough set shrinking classifier in MRS can recognize 21 patients in train set (8 patients have metastasis and the other 13 patients are free), and remaining 57 training patients are all identified as unknown type. When constructing this rough set classifier, only 9 rules extracted from the dynamic reduct {diameter (mm), ERp, PRp} are retained (see Tables 1 and 2). We note that none of the 9 rules involves the ERp parameter although the reduct does, which suggests that ERp is a valuable clinical parameter to be a distinct indicator for phenotype but not a factor leading difficulty to classification [12]. These 9 rules can further be divided into two different groups. One group has 5 rules referring to the clinical parameter of tumor diameter and the other group has 4 rules referring to the parameter PRp. In both groups, the class type (metastasis or free) is not consistently related to the parameters: when tumor diameter is 10 or 12 cm, the patients are free of metastasis; while when it is fewer as 9 cm or more as 27/35 cm, those patients suffer disease recurrence. The similar phenomenon is also observed in the rules about PRp parameter: when PRp equals 20, the patients are found to be poor prognosis; while other patients with PRp being fewer as 10 or more as 30/40 actually belong to good prognosis.

Next, from the comparison results shown in Table 3, we can see that the prediction performances of MRS are prior to the compared traditional clinical methods or even newly genetic methods. MRS can correctly predict all of 12 poor prognosis patients and 6 of 7 good prognosis patients. We emphasize that the incorrectly classified patient is just the so-called singular patient, which should be recognized by the first layer of mixture classifier yet not. The reason is that: the original train set with 78 patients possibly contains more than one group of relative singular patients, while the currently realized mixture classifier construction process only run one time so that only one group of singular patients is identified. If we wish to separate all potential groups of singular patients, more than one rough set shrinking classifiers should be needed (in terms of mixture classification methodology, it means that the mixture classification function should have more than two component functions). Since it is difficult to train multiple rough set shrinking classifiers in one mixture classifier under the disjointness constraint, it is unavoidable that some singular patients cannot be recognized by only one rough set shrinking classifier in our experiment.

Moreover, the C-SVM trained on remaining patients can also be directly used to predict the original whole 19 patients in prediction set. From the result within brackets in the second row in Table 3, we can see that the C-SVM even obtains the same prediction accuracy as the mixture classifier (although it has worse performance on original train set). This result strongly supports our assumption behind the mixture classification model: one group of singular samples in train set is relative to some singular samples in prediction set. It is interesting that the C-SVM classifier constructed on the whole train set even has worse classification ability than that one constructed on a part of original train set, which confirms again that it is possible to obtain high prediction accuracy by omitting the singular samples in train set or disregarding the singular samples in prediction set. Of course, the singular samples recognition (or the outlier classification) is an important problem in classification study, and it is still an open question that how to handle well with such singular samples. One of our research purposes is to make full use of clinical data in decision/prognosis system. As we have expected, the proposed mixture classifier successfully recognizes some singular samples

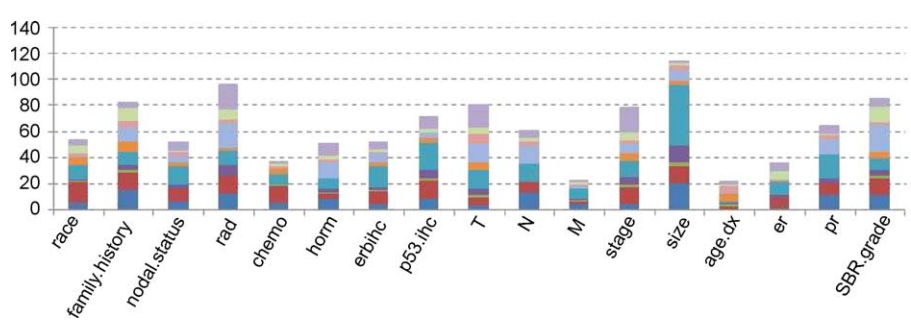

Fig. 4. The occurrence frequencies of 17 clinical parameters in rules of rough set classifiers.

and learns them by rules, while the other samples are learned by SVM to avoid unknown class type prediction and keep the high classification ability.

There are also some interesting results in the experiments on ER positive or negative patients. For ER+ cohort, its train set contains 56 patients from original train set (21 metastasis and 35 free) and its prediction set contains 12 patients from original prediction set (6 metastasis and 6 free). So for ER-cohort, its train set contains 22 patients (13 metastasis and 9 free) and its prediction set contains 7 patients (6 metastasis and 1 free). The mixture classifiers for such two groups of patients also obtain higher prognostic ability than most reported ones, which are listed at the first row in Tables 4 and 5. However, comparing with C-SVM classifiers, the AUC values of such two mixture classifiers both have an obvious decrease (see last two rows in Tables 4 and 5). This is possibly because of several factors: the number of remaining patients after the construction of rough set shrinking classifier is not sufficient for building a perfect SVM classifier; or the used criterion of rules shrinking does not fit well to the subset of ER+/ER−; or the positive and negative patients display different unbalance in those three groups (46:51 in original 97 patients group, 27:41 in ER+ group, and 19:10 in ER- group). Nevertheless, Edén has ever found that the prognosis of patients with ER− is difficult than those with ER+ in his research [12], which is also supported by our experiment results in Tables 4 and 5.

Considering the bias of clinical prognosis, there are only two kinds of mistakes: metastasis to free, or free to metastasis. The first error will lead physicians to postpone the cure for those potential serious patients in due course, and the second fault will let patients suffer unnecessary toxicity from additional therapy. Further investigating on the results, we found that 3 patients with metastasis are incorrectly judged to be free whereas no metastasis patients are misclassified in ER− model, which suggests that the ER+ patients with metastasis might suffer risk of being treated to be free when the proposed combined clinical markers are used in breast cancer prognosis.

5.2. Significance of clinical parameters of potential MRSs on BRC-2 dataset

The similar experiments on BRC-2 dataset, with results listed in Table 6, further confirm that MRS can obtain higher accuracy than the representative methods in Weka. Among those methods, J48 and RandomForest are rule-based methods, while NaiveBayes and SVM are non-rule-based classifiers. From the table we can see that, the rule-based methods are inferior to the non-rule-based method in terms of their ACC and MCC performances, greatly supporting that non-rules-based classifiers always have stronger class-recognition capability than rule-based ones. Although AttributeSelectedClassifier seems to gain benefits from the meta-analysis, it is still less effective than the parameter-optimized SVM. In all, when applied to cancer prognosis, rules-based methods offer readable reasons for therapy decisions, but they can not promise the high accuracies; while non-rule-based methods can guarantee their performances but it is hard for them to extract easily understanding rules. Although ensemble strategy will be helpful for promoting the robustness of classifiers, the ensemble classifiers are always significantly affected by the voting schemes. Moreover, if the ensembles come from the same kind of members (for example, all ensembles are rules-based classifiers), they are not expected to obtain the greatest progress on prediction. Therefore, in this paper, we propose MRS to assemble rules-based and non-rules-based methods simultaneously by introducing two-layer structure without voting. The comparison result also confirms that our idea of mixture classification is reliable. Especially, MRS greatly promotes the sensitivity indicator, which means that fewer metastasis-free patients would be misjudged as metastasis and have to receive unnecessary adjuvant therapy.

On the one hand, comparing with BRC-1, BRC-2 has nearly twice samples and clinical parameters, so that more MRS classifiers with similar accuracies would be constructed. On the other hand, the cross-validation evaluation method is used on BRC-2 for it has no independent set. Therefore, many different MRS classifiers are built on portion of original whole data. This fact means that different best rough set shrinking classifiers will be observed in different optimal MRSs, and the key clinical parameters involving the rules of those rough set shrinking classifiers will also be different. Thus we raise an interesting question that what clinical parameters have most probabilities to reflect the singular samples disturbbing the cancer prognosis. In order to answer it, we tried to count the occurrence frequencies of all 17 clinical parameters in the rules of all rough set shrinking classifiers, and found that size (diameter) and Pr (PRp) are two most significant factors in this experiment, which are also discovered in the first experiment, and actually it has been pointed out that the tumor size is an unreliable predictor of prognosis [29]; the third significant parameter is SBR grade (Scarff-Bloom-Richardson grade), which has been demonstrated to have greatly efficient prognosis value especially when the nodal status of patients is negative or unknown [30], while BRC-2 contains node-positive and node-negative samples; and the fourth significant parameter is family.history as shown in Fig. 4, which has been just reported to be closely related with the survival in breast cancer [31]. Thus, the rules-based rough set classifiers constructed under the strategy of "low coverage but high accuracy" are full of clinical meaning, and indeed are helpful for promoting the breast cancer prognosis accuracy in nature.

6. Conclusions

In this paper, we proposed a mixture classification method to build the clinical parameter based classifier MRS for breast prognosis. Based on meaningful clinical information, MRS can obtain similar or even better prognosis capability than traditional clinical marker based or genetic biomarker based methods, meanwhile the constructed rules-based rough set classifiers are full of clinical meaning. This suggests that the advanced classification methodology of combining rules-based and non-rules-based methods through a two-layer construction without voting would be a possible way for promoting the effectiveness of prognosis in clinical practice.

There are still some works to do for further promoting the efficiency of cancer prognosis. First, in the breast cancer therapy, identifying BARC1 and BARC2 methyl-mutations is a new direction for evaluating the cure risk of patients [32]. Second, the multi-layer mixture classification model is needed to be considered to promote the accuracy. Third, the possibility of integrating clinical data and gene expression data to form more powerful tool for the prognosis of breast cancer or other cancers should also be considered.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant Nos. 60773010 and 60970063.

References

[1] van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome in breast cancer. Nature 2002;415(6871):530–6.
[2] Jarman IH, Etchells TA, Martín JD, Lisboa PJ. An integrated framework for risk profiling of breast cancer patients following surgery. Artificial Intelligence in Medicine 2008;42(3):165–88.

[3] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531–7.
[4] Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, et al. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. Proceedings of the National Academy of Sciences of the United States of America 2001;98(24):13790–5.
[5] Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 2002;1(2):203–9.
[6] van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene expression signature as a predictor of survival in breast cancer. The New England Journal of Medicine 2002;347(25):1999–2009.
[7] Spentzos D, Levine DA, Ramoni MF, Joseph M, Gu X, Boyd J, et al. Gene expression signature with independent prognostic significance in epithelial ovarian cancer. Journal of Clinical Oncology 2004;22(23):4700–10.
[8] Spentzos D, Levine DA, Kolia S, Out H, Boyd J, Libermann TA, et al. Unique gene expression profile based on pathologic response in epithelial ovarian cancer. Journal of Clinical Oncology 2005;23(31):7911–8.
[9] Li L. Survival prediction of diffuse large-b-cell lymphoma based on both clinical and gene expression information. Bioinformatics 2006;22(4):466–71.
[10] Dunkler D, Michiels S, Schemper M. Gene expression profiling: does it add predictive accuracy to clinical characteristics in cancer prognosis? European Journal of Cancer 2007;43(4):745–51.
[11] Gevaert O, De Smet F, Timmerman D, Moreau Y, De Moor B. Predicting the prognosis of breast cancer by integrating clinical and micro-array data with Bayesian networks. Bioinformatics 2006;22(14):e184–90.
[12] Edén P, Ritz C, Rose C, Fernö M, Peterson C. Good old clinical markers have similar power in breast cancer prognosis as micro-array gene expression profilers. European Journal of Cancer 2004;40(12):1837–41.
[13] West D, Mangiameli P, Rampal R, West V. Ensemble strategies for a medical diagnostic decision support system: a breast cancer diagnosis application. European Journal of Operational Research 2005;162(2):532–51.
[14] Geurts P, Fillet M, de Seny D, Meuwis MA, Malaise M, Merville MP, et al. Proteomic mass spectra classification using decision tree based ensemble methods. Bioinformatics 2005;21(14):3138–45.
[15] Peng Y. A novel ensemble machine learning for robust microarray data classification. Computers in Biology and Medicine 2006;36(6):553–73.
[16] Ge G, Wong GW. Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. BMC Bioinformatics 2008;9:275.
[17] Koziol JA, Feng AC, Jia Z, Wang Y, Goodison S, McClelland M, et al. The wisdom of the commons: ensemble tree classifiers for prostate cancer prognosis. Bioinformatics 2009;25(1):54–60.
[18] Ahr A, Holtrich U, Solbach C, Scharl A, Strebhardt K, Karn T, et al. Molecular classification of breast cancer patients by gene expression profiling. The Journal of Pathology 2001;195(3):312–20.
[19] Perreard L, Fan C, Quackenbush JF, Mullins M, Gauthier NP, Nelson E, et al. Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. Breast Cancer Research 2006;8(2):R23.
[20] Hirose S, Shimizu K, Kanai S, Kuroda Y, Noguchi T. POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. Bioinformatics 2007;23(16):2046–53.
[21] Witten IH, Frank E. Data mining: practical machine learning tools and techniques with java implementations. San Francisco, CA: Morgan Kaufmann; 1999.
[22] Smith DD, Sætrom Jr P, Lundberg OS, Rivas C, Glackin GEC, et al. Meta-analysis of breast cancer microarray studies in conjunction with conserved cis-elements suggest patterns for coordinate regulation. BMC Bioinformatics 2008;9:63.
[23] Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, et al. Meta-analysis of gene-expression profiles in breast cancer: toward a unified understanding of breast cancer sub-typing and prognosis signatures. Breast Cancer Research 2008;10:R65.
[24] Bazan JG, Szczuka MS, Wroblewski J. A new version of rough set exploration system. In: TSCTC'02: proceedings of the third international conference on rough sets and current trends in computing. London, UK: Springer-Verlag; 2002. p. 397–404.
[25] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. URL http://www.csie.ntu.edu.tw/~cjlin/libsvm [accessed: 29/06/2009]; 2001.
[26] Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell 2006;10(6):529–41.
[27] Elfilali A, Lair S, Verbeke C, La Rosa P, Radvanyi F, Barillot E. ITTACA: a new database for integrated tumor transcriptome array and clinical data analysis. Nucleic Acids Research 2006;34(Database issue):D613–6.
[28] Gruvberger S, Ringnér M, Chen Y, Panavally S, Saal LH, Borg A, et al. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. Cancer Research 2001;61(16):5979–84.
[29] Foulkes WD, Grainge MJ, Rakha EA, Green AR, Ellis IO. Tumor size is an unreliable predictor of prognosis in basal-like breast cancers and does not correlate closely with lymph node status. Breast Cancer Research and Treatment 2009;117(1):199–204.
[30] Le Doussal V, Tubiana-Hulin M, Friedman S, Hacene K, Spyratos F, Brunet M. Prognostic value of histologic grade nuclear components of Scarff–Bloom–Richardson (SBR), an improved score modification based on a multivariate analysis of 1262 invasive ductal breast carcinomas. Cancer 1989;64(9):1914–21.
[31] Hemminki K, Ji J, Försti A, Sundquist J, Lenner P. Survival in breast cancer is familial. Breast Cancer Research and Treatment 2008;110(1):177–82.
[32] Rennert G, Bisland-Naggan S, Barnett-Griness O, Bar-Joseph N, Zhang S, Rennert HS, et al. Clinical outcomes of breast cancer in carriers of BRCA1 and BRCA2 mutations. The New England Journal of Medicine 2007;357(2):115–23.