Computers in Biology and Medicine 40 (2010) 179–189

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction

Shu-Lin Wang [a,b], Xueling Li [b], Shanwen Zhang [b], Jie Gui [b,c], De-Shuang Huang [b,*]

[a] School of Computer and Communication, Hunan University, Changsha, Hunan 410082, China
[b] Intelligent Computation Lab, Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui 230031, China
[c] Department of Automation, University of Science and Technology of China, Hefei, Anhui 230026, China

ARTICLE INFO

Article history:
Received 2 October 2008
Accepted 29 November 2009

Keywords:
Biological data mining
Gene expression profiles
Gene selection
Neighborhood rough set model
Probabilistic neural network ensemble
Tumor classification

ABSTRACT

Since Golub applied gene expression profiles (GEP) to the molecular classification of tumor subtypes for more accurately and reliably clinical diagnosis, a number of studies on GEP-based tumor classification have been done. However, the challenges from high dimension and small sample size of tumor dataset still exist. This paper presents a new tumor classification approach based on an ensemble of probabilistic neural network (PNN) and neighborhood rough set model based gene reduction. Informative genes were initially selected by gene ranking based on an iterative search margin algorithm and then were further refined by gene reduction to select many minimum gene subsets. Finally, the candidate base PNN classifiers trained by each of the selected gene subsets were integrated by majority voting strategy to construct an ensemble classifier. Experiments on tumor datasets showed that this approach can obtain both high and stable classification performance, which is not too sensitive to the number of initially selected genes and competitive to most existing methods. Additionally, the classification results can be cross-verified in a single biomedical experiment by the selected gene subsets, and biologically experimental results also proved that the genes included in the selected gene subsets are functionally related to carcinogenesis, indicating that the performance obtained by the proposed method is convincing.

1. Introduction

Tumor is identified as systematic biology diseases [1]. So far the mechanism of tumor development is not thoroughly known yet. Since tumor treatment of patients of later stage cancers is often not therapeutically effective, medical experts agree that early diagnosis of tumor is of great benefit to the successful therapies of tumor. However, it is difficult for traditional tumor mass detection techniques, such as X-ray imaging, to conduct early detection of tumor. In recent 10 years, gene expression profiles (GEP) based molecular diagnosis of tumor have attracted a great number of medical researchers and computer scientists for the goal of realizing precise and early tumor diagnosis [2–6]. However, the curse of dimensionality caused by high dimensionality and small sample size of tumor dataset seriously challenges the tumor classification. So how to select important gene subsets from thousands of genes in GEP dataset to drastically reduce the dimensionality of tumor dataset is the first key step to address this problem. Usually, the prediction performance of the selected gene subsets is evaluated by a classifier. The commonly used classifiers including support vector machines (SVM) [7–10], *k*-nearest neighbor (*k*-NN) [11,12], C4.5 [13], artificial neural networks (ANN) [14,15], self-organizing map (SOM) [16], self-organizing tree algorithm (SOTA) [17], and probabilistic neural networks (PNN) [18–20] have been extensively applied to the molecular classification of tumor subtypes for more accurately and reliably clinical diagnosis. From those experimental results, we could conclude that gene selection such as selecting informative genes by using regulation probability [21] and by using independent component analysis [22] plays an important role in tumor classification.

Finding minimum tumor-related gene subsets can really improve the predictive performance of classification model because too many redundant or irrelevant genes might degrade the classification accuracy [23]. In addition to removing noise in GEP, the selected gene subsets also have important biomedical meanings and may be applied to the discovery of drug targets. Generally speaking, gene selection methods are categorized into two groups [24]. One is Wrapper methods which combine gene selection with a classifier, and another is Filter methods in which the procedure of gene selection is independent of classifiers. In most cases, Wrapper methods is superior to Filter methods in

* Corresponding author. Tel./fax: +86 551 5592751.
E-mail address: dshuang@iim.ac.cn (D.-S. Huang).

0010-4825/$ - see front matter © 2009 Published by Elsevier Ltd.
doi:10.1016/j.compbiomed.2009.11.014

improving classification accuracy [25]. However, Wrapper methods by adopting different classifiers usually obtain different optimal gene subsets, which indicates that the Wrapper methods would be unstable in gene selection to some extent because the obtained accuracy is sensitive to the selected gene subsets. Another drawback is their high computational time. These are intrinsic drawbacks for most of the existing Wrapper methods when facing the curse of dimensionality and a variety of uncertainties in tumor dataset (the gathering process of microarray data including fabrication, hybridization and image processing always adds various sources of noise) [26,27]. To address these problems, traditional intelligent methods are apt to overfitting in classifying tumor dataset due to the lack of training sample set [28]. In fact, there are numerously optimal gene subsets with very high classification accuracy in tumor dataset [29,30], which is mainly caused by gene co-expression and the function similarity of many genes, so how to obtain convincingly classification accuracy from these optimal gene subsets is still an important problem.

Solutions to the above problem include various ensemble schemes [31–35]. These studies suggested that ensemble machine learning or classifiers consistently perform better [13] in that a powerful classifier can be constructed by the ensemble of many base classifiers even though these base classifiers are weak in making decisions [36]. For example, Peng [27] proposed a robust ensemble approach to tumor classification by generating a pool of candidate base classifiers based on gene sub-sampling and then selecting a set of appropriate base classifiers to construct a high performance classification committee based on classifier clustering. Both theoretical and experimental studies have shown that the integrating of a set of diverse and accurate base classifiers would lead to a powerful ensemble classifier, where the diversity of base classifiers is prerequisite to the powerful ensemble classifier that outperforms each base classifier [37], because combining a set of same classifiers will not intuitively generates any improvement. However, most of the conventional ensemble methods employed to tumor classification such as re-sampling methods based on samples or gene re-sampling are so random that their biological meanings are difficult to interpret. Therefore, the diversity and accuracy of base classifiers should be considered simultaneously in designing an ensemble classifier. In this study, we propose a novel ensemble method which combines base PNN classifiers with neighborhood rough set model based gene reduction. Experiments on three well-known tumor datasets show that the proposed methods not only have higher classification accuracy rate but also are more stable in classification performance.

The remainder of this paper is organized as follows. In Section 2, we first introduced the neighborhood rough set model for gene reduction, the framework of PNN ensemble algorithm and two gene pre-selection methods: an iterative search margin based algorithm and a weighted feature score criterion. Section 3 described our four experimental methods and provided their experimental results on three well-known tumor datasets and the biomedical interpretation of some selected genes. Comparison with other related works were also roughly performed in this section. Finally, Section 4 presented the conclusions.

2. Methods

2.1. Neighborhood rough set model

How to generate diverse base classifiers is a critical problem in ensemble machine learning. In our ensemble method, diverse base classifiers were produced by diverse gene subsets obtained by using gene reduction based on neighborhood rough set model (NRSM) [38,39]. The principle of NRSM was briefly introduced as follows.

Let $G=\{g_1,\dots g_n\}$ be a set of genes and $S=\{s_1,\dots s_m\}$ be a set of samples. The corresponding gene expression matrix can be represented as $X=(x_{i,j})_{m\times n}$, where $x_{i,j}$ is the expression level of gene g_j in sample s_i, and usually $n\gg m$, and here m is the number of samples, and n is the number of genes. The matrix X is composed of m row vectors $s_i\in R^n, i=1,2,\dots,m$. Each vector s_i in the gene expression matrix may be regarded as a point in n-dimensional space, and each of the n columns consists of an m-element expression vector for a single gene.

Let $NDT=\langle S,G\cup D,V,f\rangle$ be a neighborhood decision table, where $S=\{s_1,\dots,s_m\}$ is a nonempty sample set called sample space, and $G=\{g_1,\dots,g_n\}$ is a nonempty set of genes called condition attributes, $D=\{l_1,l_2,\dots,l_c\}$ is an output variable called decision attribute which denotes tumor subclasses, V_a is a value domain of attribute $a\in G\cup D$, f is an information function $f: S\times(G\cup D)\to V$, where

$$V=\bigcup_{a\in G\cup D} V_a$$

Given $\forall s_i\in S$ and $B\subseteq G$, the neighborhood $\delta_B(s_i)$ of s_i in the subspace B is defined as

$$\delta_B(s_i)=\{s_j|s_j\in S,\Delta_B(s_i,s_j)\le\delta\}$$

where δ is a threshold value, and $\Delta_B(s_i,s_j)$ is a metric function in subspace B. There are three common metric functions that are widely used. Let s_1 and s_2 be two samples in n-dimensional space $G=\{g_1,\dots,g_n\}$. $f(s,g_i)$ denotes the value of g_i in sample s. Then Minkowsky distance is defined as:

$$\Delta_p(s_1,s_2)=\Big(\sum_{i=1}^{n}|f(s_1,g_i)-f(s_2,g_i)|^p\Big)^{1/p}$$

where (1) when $p=1$, it is called Manhattan distance Δ_1; (2) when $p=2$, it is Euclidean distance Δ_2; (3) when $p=\infty$, it is Chebychev distance.

Given a neighborhood decision table NDT, $X_1,X_2,\dots,X_c$ are the sample subsets with decisions l_1 to l_c, $\delta_B(x_i)$ is the neighborhood information granules including x_i and is generated by gene subset $B\subset G$, then the lower and upper approximations of the decision D with respect to gene subset B are respectively defined as

$$Lower(D,B)=\bigcup_{i=1}^{c} Lower(X_i,B)$$

$$Upper(D,B)=\bigcup_{i=1}^{c} Upper(X_i,B)$$

where $Lower(X,B)=\{x_i|\delta_B(x_i)\subseteq X,x_i\in S\}$ is the lower approximations of the sample subset X with respect to gene subset B, and called positive region denoted by $Pos(D,B)$, and $Upper(X,B)=\{x_i|\delta_B(x_i)\cap X\neq\phi,x_i\in S\}$ is the upper approximations.

The decision boundary region of D to B is defined as

$$BN(D,B)=Upper(D,B)-Lower(D,B).$$

The dependency degree of D to B is defined as the ratio of consistent objects $\gamma(D,B)=Card(Pos(D,B))/Card(S)$. Here $\gamma(D,\phi)=0$ is defined, and $Card(S)$ denotes the cardinal number of sample set S. Let $a\in B$, then the significance of the gene a is defined as

$$SIG(a,D,B)=\gamma(D,B\cup a)-\gamma(D,B).$$

Hu [38] designed a greedy search algorithm named forward attribute reduction based on neighborhood model (FARNeM). We employed this algorithm to generate diverse optimal gene subsets by taking different δ values.