Applied Soft Computing 66 (2018) 50–59

Contents lists available at ScienceDirect

Applied Soft Computing

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

A novel effective diagnosis model based on optimized least squares support machine for gene microarray

Xinteng Gao, Xinggao Liu*

National Engineering Research Center of Industrial Automation, College of Control Science & Engineering, Zhejiang University, Hangzhou 310027, PR China

ARTICLE INFO

Article history:
Received 30 December 2016
Received in revised form 8 January 2018
Accepted 7 February 2018
Available online 12 February 2018

Keywords:
Microarray data
SVMRFE
PSOFOA
LSSVM
Disease diagnosis

ABSTRACT

Classification for cancer diagnosis is a significant topic in bioinformatics. Microarray data generates a big challenge for the diagnosis prediction due to tens of thousands of genes but dozens of samples. The feature dimension needs to be reduced for better diagnosis prediction. Moreover, the parameters of classifiers required to be optimized have a significant impact on the classifiers' performance. This paper presents a novel effective approach for diagnosis prediction. First, the normalization procedure for the raw data is carried out. Then, considering both the complexity of dimension of the microarray data and the specific classifier, support vector machine method based on recursive feature elimination (SVMRFE) is applied for selecting optimal gene subset. Finally, a hybrid method based on fruit fly optimization and particle swarm optimization (PSOFOA) is proposed to optimizing the classifier (LSSVM) parameters. An integrated work of disease diagnosis is obtained. By this means, the accuracy of 100% is achieved with only 4 features (genes), which is the best result compared to all published papers. The proposed method is also compared with some specific methods The result illustrates that the proposed method achieves a best prediction performance.

1. Introduction

A DNA microarray (also commonly known as DNA chip or biochip) is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure the expression of large numbers of genes simultaneously. Thanks to the DNA microarray, researchers can detect tens of thousands of genes by only one experiment [1,2]. This convenience helps the researchers gain an insight into disease at the gene level with much less money and time consumption than before. However, restricted by the objective conditions, DNA microarray data has the typical characteristics of high dimension and small samples [3]. There are usually only one hundred or even dozens of samples in an experiment. Meanwhile, thousands or even tens of thousands of genes are measured in a sample, among which there are obviously a lot of invalid or redundant ones [4].

In this paper, acute leukemia is chosen as a test case. Classification of acute leukemia began with the observation of variability in clinical outcome and subtle differences in nuclear morphology [5]. Before the use of microarray data, no single test was sufficient to establish the diagnosis. Rather, an experienced hematopathologist's interpretation of the tumor's morphology, histochemistry, immunophenotyping, and cytogenetic analysis were needed. Even by that mean, leukemia classification was imperfect and errors did occur very often [6]. Such problem is possibly solved by the use of gene microarray data scientifically [7]. Among these published literatures with this dataset, some achieved a good diagnosis accuracy with a large number of features, some got a relatively poorer result in order to reduce the feature dimension as far as possible [8]. In reference [7], researchers firstly sorted genes by their degree of correlation. Neighborhood analysis showed that about 1000 genes were highly correlated with this class distinction work. Secondly, known samples were used to create a class predictor. Each of the informative genes cast a "weighted vote" for one of the classes. Then, the votes were summed to determine the winning class [7]. Such classification algorithm for the AML/ALL dataset was enlightened. However, preprocessing of the raw data wasn't carried out and the accuracy was barely acceptable. A popular step of preprocessing such data was introduced in reference [9] to make the whole work more reliable.

On how to effectively reduce the dimension and the size of the data, unearth decisive genes of disease, former researchers made a lot of work [10]. There are three classical techniques of feature selection: filters, wrappers and embedded methods [11]. Filter

* Corresponding author.
E-mail address: lxg@zju.edu.cn (X. Liu).

https://doi.org/10.1016/j.asoc.2018.02.009
1568-4946/

methods are based on information theory [12]. The advantage of filter methods is low computing complexity. Shipp et al. applied SNR method to the diffuse large B cell lymphoma (DLBCL) dataset, where 30 features were used to classify DLBCL and FL (follicular lymphoma) and the accuracy of 91% was achieved [13]. Li et al. proposed a feature selection method based on discrete wavelet transform (DWT) and applied it to the colon cancer dataset, achieving the accuracy of 93.55% [14]. Filter methods cannot, however, give effective evaluation between feature subsets. Wrapper methods treat different feature subsets as training data of machine learning algorithm. The model's prediction ability is used to evaluate the performance of these feature subsets for the feature selection [15]. This kind of method treats the classifier as a black box in the feature selection process, and the subset with highest classification accuracy is considered as optimal one [16]. Chu et al. used the recursive circulation of fuzzy neural network (FNN) in the microarray data and achieved 100% accuracy with 8 genes [17] while Tibshirani et al. achieved the same accuracy with 43 genes [18]. Wrapper methods have the risk of over fitting. Furthermore, as the number of features grows, the space of feature subsets grows exponentially. An intermediate solution of the drawbacks of filters and wrappers for researchers is the embedded method, which uses the core of the classifier to establish a criterion for ranking features [19]. Tan et al. [20] applied SVMRFE, *t*-test [21], entropy based methods to colon cancer dataset and prostate cancer dataset. It seemed that there was not an optimal classifier for all datasets [22–24].

Among these methods, SVM is a small sample learning method with solid theoretical basis. It basically does not involve the probability measurement; essentially avoid the traditional process from induction to deduction, simplifying the classification and regression problems. At the same time, the complexity of its calculation depends on the number of support vectors, not the dimension of the sample space, which in some sense avoids the "dimensionality disaster". In recent years, least squares support vector machine (LSSVM), an extension of SVM, has been attracting people's attention. LSSVM simplifies the solution of SVM optimization problem, improving the speed of solution and is more suitable for general application. Now it is also applied to the large-scale problem such as gene microarray data analysis and the pattern recognition problem such as feature selection [25]. Many researches have made outstanding achievements [17,26]. In this paper, SVM and LSSVM are both introduced and used in the diagnosis model. SVM is used for initial filtering and sorting the features. LSSVM is used as the classifier in the diagnosis model. Besides, FOA and PSO are used as the optimization algorithms in this paper due to their good performance.

In this paper, an integrated hybrid diagnosis model is proposed to achieve the accuracy of 100% with only 4 features. A filter method (F-statistics) is used to quickly rule out part of the noise data. In this step, top 200 genes are selected. Then the embedded method (SVMRFE) is used for ranking genes to choose the optimal feature subset. In the step of parameters optimization, 2 swarm intelligence algorithms are combined for searching global optimal parameters.

This paper is organized as follows. Section 2 provides the theoretic descriptions of the SVMRFE, PSO, FOA and LSSVM. Case study is presented in Section 3. Section 4 contains the results and discussions of the case study. Finally, Section 5 concludes this paper.

2. Methods

2.1. F-statistics

After the step of SVMRFE, the classifier will be retrained and the remaining features will be reevaluated. The computational cost of SVMRFE algorithm is proportional to the number of features. In order to reduce the dimension of the problem, F-statistics is used as a filter method before the step of SVMRFE [27].

F-statistics is defined as

$$R(j) = \frac{\left(\sum_{i=1}^{m}\sum_{k=1}^{2} 1_{(y_i=k)}(\overline{x_{kj}} - \overline{x_j})^2\right)}{\left(\sum_{i=1}^{m}\sum_{k=1}^{2} 1_{(y_i=k)}(x_{ij} - \overline{x_{kj}})^2\right)}, 1 \le j \le n \quad (1)$$

where

$\overline{x_j}$: The mean value of gene j in all samples.

$\overline{x_{kj}}$ The mean value of gene j in class k.

x_{ij}: The value of gene j in sample i.

1_{Ω}: Indicator function of event Ω, when Ω is true, $1_{\Omega} = 1$, otherwise, $1_{\Omega} = 0$.

The genes with higher F scores are considered to be more important. In this paper, this method is used to choose the 200 most important genes with high F score relatively.

2.2. Least squares support vector machine (LSSVM)

The support vector machine method (SVM) introduced by Vapnik et al. [28] in 1990s is an excellent tool for classification and regression, especially for the cases with very few high-dimensional samples [29]. The LSSVM (least square support vector machine) is an extension of SVM, which applies the linear least squares criteria as loss function instead of the quadratic programming problem [30].

Suppose that there is a samples set $\{x_i, y_i\}_{i=1}^{n}$ with the sample x_i and the corresponding sample class label y_i. The quadratic norm of error ξ_i is taken as the loss function of LSSVM. The following optimization problem is considered:

$$\min_{w,b,\xi} J(w, \xi) = \frac{1}{2}w^T w + \gamma\frac{1}{2}\sum_{i=1}^{n}\xi_i^2 \quad (2)$$

which is subject to the equality constraints:

$$y_i = w^T\varphi(x_i) + b + \xi_i^2, i = 1, 2, \ldots, n \quad (3)$$

The least square support vector machine model is as follows:

$$f(x) = w^T\varphi(x) + b \quad (4)$$

J is the objective function, $\frac{1}{2}w^T w$ is treated as the flatness measurement function. γ is the punishment factor taking consideration of tradeoff between the training error and the model flatness, and $\xi_i = \alpha_i/\gamma$ is the error variable; the nonlinear mapping φ maps the input training data into a high-dimensional feature space, in which a linear regression problem is obtained and to be solved; b is the bias, and w is a weight vector of the same dimension as the feature space [31].

The Lagrangian function $L(\cdot)$ can be constructed by:

$$L(\omega, b, \xi, \alpha) = \frac{1}{2}w^T w + \gamma\frac{1}{2}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i(\xi_i - y_i + w^T\varphi(x_i) + b) \quad (5)$$