Genomics Data 5 (2015) 46–50

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Interval-value Based Particle Swarm Optimization algorithm for cancer-type specific gene selection and sample classification

D. Ramyachitra *, M. Sofia, P. Manikandan

Department of Computer Science, Bharathiar University, Coimbatore 641046, India

ARTICLE INFO

Article history:
Received 8 April 2015
Received in revised form 27 April 2015
Accepted 29 April 2015
Available online 23 May 2015

Keywords:
Microarray
Gene selection
Tissue sample classification
Particle swarm optimization
Interval-value classification
Interval-value based Particle Swarm Optimization classification

ABSTRACT

Microarray technology allows simultaneous measurement of the expression levels of thousands of genes within a biological tissue sample. The fundamental power of microarrays lies within the ability to conduct parallel surveys of gene expression using microarray data. The classification of tissue samples based on gene expression data is an important problem in medical diagnosis of diseases such as cancer. In gene expression data, the number of genes is usually very high compared to the number of data samples. Thus the difficulty that lies with data are of high dimensionality and the sample size is small. This research work addresses the problem by classifying resultant dataset using the existing algorithms such as Support Vector Machine (SVM), K-nearest neighbor (KNN), Interval Valued Classification (IVC) and the improvised Interval Value based Particle Swarm Optimization (IVPSO) algorithm. Thus the results show that the IVPSO algorithm outperformed compared with other algorithms under several performance evaluation functions.

1. Introduction

Cancer classification using gene expression data usually relies on traditional supervised learning techniques, in which only labeled data (i.e., data from a sample with clinical follow-up) can be exploited for learning, while unlabeled data (i.e., data from a sample without clinical follow-up) are disregarded [1]. Recent research in the area of cancer diagnosis suggests that unlabeled data in addition to the small number of labeled data can produce significant improvement in terms of accuracy by using a technique called semi supervised learning. Indeed, semi supervised learning has proved to be effective in solving different biological problems including protein classification, prediction of transcription factor–gene interaction and gene-expression based cancer subtype discovery. Microarray technology allows simultaneous measurement of the expression levels of thousands of genes within a biological tissue sample. An important application of gene expression is to classify samples according to their gene expression profiles, such as the diagnosis or the classification of different types or subtypes of cancer [2,3]. Different classification methods from statistical and machine learning have been applied to the classification of cancer. However, high dimensionality and possibly a small number of noisy samples pose great challenges to the existing methods. The main approach to this problem was based on the existing algorithms to analyze gene expression data. Most of the classifiers involve complex models containing numerous genes. This has limited the interpretability of the classifiers and this lack of interpretability hampers the acceptance of diagnostic tools. Classification models based on numerous genes can also be more difficult to transfer to other assay platforms, which may be more suitable for clinical application. Several researchers pointed out that the classifiers might be developed to contain a small number of genes that provide classification accuracy comparable to that achieved by models that are more complex [4]. Moreover, some more complex algorithms based on numerous genes for classification often overfit the data [5].

Prior to classification, a variety of gene selection strategies have been used. The aim of gene selection is to select a small subset of genes from a larger pool [6,7]. Gene selection methods are classified into three types: (1) filter methods, (2) wrapper methods and (3) embedded methods. Filter methods evaluate a subset of genes by looking at the intrinsic characteristics of data with respect to class labels, while wrapper methods evaluate the goodness of a gene subset by the accuracy of its learning or classification. Embedded methods are generally referred to as algorithms, where gene selection is embedded in the construction of the classifier. In the gene selection process, an optimal feature subset is always relative to a certain criterion. Every criterion measures the discriminating ability of a gene or a subset of genes to distinguish different class labels. To measure the gene–class relevance, different statistical and theoretical measures such as the t-test, entropy and mutual information are typically used, and different metrics including the Euclidean distance and correlation coefficient are employed to calculate the gene–gene redundancy [11,15].

* Corresponding author. Tel.: +91 99 943 74 370.
E-mail address: jaichitra1@yahoo.co.in (D. Ramyachitra).

http://dx.doi.org/10.1016/j.gdata.2015.04.027

In filters, the characteristics in the feature selection are uncorrelated to those of the learning methods, therefore they have a better generalization property [1]. The filters, wrapper and embedded are then analyzed to identify the most frequently appearing genes which would correspond to the most predictive genes [2]. The Genetic Algorithm combined with a Support Vector Machine classifier is used for selecting predictive genes and for final gene selection and classification. The analysis of gene expression data is to identify the sets of genes as classification or diagnosis platforms. Machine learning techniques, such as artificial neural networks (ANNs), present a more flexible 'model-free' approach for classification and frequently yield good results [6]. The advantage of selecting a combination of genes with small redundancy, favors the selection of mutually uncorrelated genes. The selected set of paired genes was used as a new feature set for the classification.

In wrapper type methods, feature selection is "wrapped" around a learning method and a feature is directly judged by the estimated accuracy of the learning method [11]. One can often obtain a set with a very small number of non-redundant features, which gives high accuracy, because the characteristics of the features match well with the characteristics of the learning method [14]. Wrapper methods can use different performance metrics and objective functions. And also the wrapper methods select the "minimum" subset of features that provides the highest sensitivity. Embedded methods differ from other feature selection methods in the way that feature selection and learning interact [14]. In contrast to filter and wrapper approaches, in embedded methods the learning part and the feature selection part cannot be separated — the structure of the class of functions under consideration plays a crucial role [22].

2. Experiments

In this section, we evaluate the discriminative performance of our selected gene set on different classifiers. We also compare the performance of our proposed classification method to a wide range of standard classifiers: Support Vector Machine (SVM), K Nearest Neighbor (KNN), Particle Swarm Optimization (PSO) and Interval Value Classification (IVC). A set of experiments is conducted on the dataset by varying the number of genes selected to receive the highest classification accuracy.

2.1. Results on the leukemia dataset

To evaluate the performance of the proposed method in practice, this research used the datasets containing gene expression profiles from patients with acute lymphoblastic leukemia (ALL) and acute myeloblastic leukemia (AML). The leukemia dataset is collected from the UCI Repository. In the leukemia dataset 72 samples are used for the training set and 32 samples are used as the testing set. This dataset have compared with the leukemia dataset that contains the ALL/AML types. The ALL portion of the dataset is derived from two cell types, B-cells and T-cells, while the AML part is split into two types as bone marrow (BM) samples and peripheral blood (PB). The correctly classified instance for the leukemia dataset is 8.0 and incorrectly classified instance is 1.0. The comparison has been done with proposed IVPSO and several existing algorithm such as SVM, KNN, IVC. It has been found that the proposed algorithm is better than the existing algorithm for classifying the leukemia datasets. Table 2.1 shows the results for the leukemia dataset and Fig. 2 shows the performance comparison of existing and proposed algorithms for the leukemia dataset.

2.2. Results on the breast cancer dataset

To further test the performance of the proposed method the breast cancer dataset is used for comparison, and it is collected from the UCI Repository. Here the dataset consists of 69 samples from human cancer cell lines. The breast cancer dataset spans nine classes and gene expression levels were measured for 769 genes. The prediction accuracy of 74.86 is reported in reference using one-versus-the rest IVC with 150 selected genes. To test the proposed algorithm on an external dataset, 43 samples are used for the training dataset while 18 samples as the testing dataset. Based on 150 genes selected and 12 genes selected by PSO, the classification accuracy report of all the compared algorithms can be predicted. The correctly classified instance for the breast cancer dataset is 7.2 and incorrectly classified instance is 2.8. Consistent with the results on the breast cancer dataset in this experiment, the proposed method also achieved the highest classification accuracy. Thus Table 2.2 shows the results for the breast cancer dataset and Fig. 3 shows the performance comparison of existing and proposed algorithms for the breast cancer dataset.

Table 2.1
Performance comparison of existing and proposed methods for the leukemia dataset.

Algorithms/performance metrics	TP rate	FP rate	Precision	Accuracy
Support Vector Machine	70.97	28.61	43.75	69.01
K-Nearest Neighbor	80.27	22.2	90.0	71.28
Interval Valued Classification	85.0	60.0	94.4	78.26
Particle Swarm Optimization	90.0	22.6	83.35	81.8
Interval Value based Particle Swarm Optimization	100	0.0	90.0	96.88

2.3. Results on the lung cancer datasets

The performance of the proposed algorithm is calculated by using the lung cancer dataset and it can be collected from the UCI Repository which consists of 61 samples from human cancer cells. In the lung cancer dataset, the class and gene expression levels were measured for 462 genes. The prediction accuracy of IVC is 70.55 with 72 instances and 32 attributes. To test the proposed algorithm, a dataset of 43 samples was used for the training dataset and 32 samples as the testing dataset. The correctly classified instance for the lung cancer dataset is 7.2 and the incorrectly classified instance is 2.8. Consistent with the results on the lung cancer dataset in this experiment, the proposed method also achieved the highest classification accuracy. Thus Table 2.3 shows the results for the lung cancer dataset and Fig. 4 shows the performance comparison of existing and proposed algorithms for the lung cancer dataset.

2.4. Results on blood cancer datasets

The performance of the proposed algorithm is also measured using the blood cancer datasets and it can be collected from the NCBI database. Blood cancer is an umbrella term for cancers that affect the bone marrow, blood and lymphatic system. In this dataset a total of 399 instances and 18 attributes were used. In this analysis, the data are based on class distribution. In 339 instances, to test the proposed algorithm a dataset of 48 samples were used for the training dataset and 36 samples as the testing dataset. The correctly classified instance is 7.8 and the incorrectly classified instances are 2.2. Consistent with the results on the blood cancer dataset with this experiment, the proposed method also achieved the highest classification accuracy. Thus Table 2.4 shows the results for the blood cancer dataset and Fig. 5 shows the

Table 2.2
Performance comparison of existing and proposed methods for breast cancer dataset.

Algorithms/performance metrics	TP rate	FP rate	Precision	Accuracy
Support Vector Machine	71.26	29.45	70.75	71.87
K Nearest Neighbor	76.8	27.24	75.95	67.29
Interval Valued Classification	80.1	25.24	75.66	74.86
Particle Swarm Optimization	82.8	20.86	79.87	84.63
Interval Value based Particle Swarm Optimization	90.16	17.17	83.9	92.24