Applied Soft Computing 48 (2016) 298–316

Contents lists available at ScienceDirect

Applied Soft Computing

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

Pipelining the ranking techniques for microarray data classification: A case study

Rasmita Dash [a,*], Dr. Bijan Bihari Misra [b]

[a] Department of Computer Sc. & Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan, University, Khandagiri Square, Bhubaneswar, 751030 Odisha, India
[b] Department of Computer Sc. & Engineering, Silicon Institute of Technology, Bhubaneswar, 751024 Odisha, India

ARTICLE INFO

Article history:
Received 7 August 2015
Received in revised form 23 March 2016
Accepted 3 July 2016
Available online 20 July 2016

Keywords:
Microarray data
Feature selection
Feature ranking technique
Classification
Statistical test

ABSTRACT

Identification of relevant genes from microarray data is an apparent need in many applications. For such identification different ranking techniques with different evaluation criterion are used, which usually assign different ranks to the same gene. As a result, different techniques identify different gene subsets, which may not be the set of significant genes. To overcome such problems, in this study pipelining the ranking techniques is suggested. In each stage of pipeline, few of the lower ranked features are eliminated and at the end a relatively good subset of feature is preserved. However, the order in which the ranking techniques are used in the pipeline is important to ensure that the significant genes are preserved in the final subset. For this experimental study, twenty four unique pipeline models are generated out of four gene ranking strategies. These pipelines are tested with seven different microarray databases to find the suitable pipeline for such task. Further the gene subset obtained is tested with four classifiers and four performance metrics are evaluated. No single pipeline dominates other pipelines in performance; therefore a grading system is applied to the results of these pipelines to find out a consistent model. The finding of grading system that a pipeline model is significant is also established by Nemenyi post-hoc hypothetical test. Performance of this pipeline model is compared with four ranking techniques, though its performance is not superior always but majority of time it yields better results and can be suggested as a consistent model. However it requires more computational time in comparison to single ranking techniques.

1. Introduction

In recent years, analysis of microarray gene expression data comes up as a challenging issue in bioinformatics research. In this context, classification of microarray samples represents a well-studied problem in statistics and machine learning, where a large number of successful methods have been suggested [1]. However, it has also been shown that commonly used baseline classifiers pose intrinsic draw-backs in achieving accurate and reproducible results. In order to obtain more robust microarray data analysis, sophisticated methods should be applied for classification and prediction of microarray data [2].

Microarray data set or gene expression data sets are organized as matrix form and is experimented on different samples. The column represents different genes in gene expression data and row represents sample measured at different time point. There are number of gene expression data analysis techniques and classification is one out of them. In classification a classifier will assign a new sample to one of the existing class.

For example in a two class problem, a sample is assigned to disease affected class (positive class) or normal class (negative class). Again in microarray data the number of genes is too high (in the range of 1000–10000) and number of sample is comparatively low (in the range of 100). This, however, poses a great challenge to traditional classification algorithms. So this high dimension increases the search space and makes the classification task more complex. So gene selection or feature selection is prerequisite for classification. For efficient high dimensional data classification Dash and Liu [3] carried out a comprehensive overview of feature selection techniques. There are different feature selection methods and they are categorised based on some criterion. For example feature selection methods may be categorised into two ways called as filter and wrapper method [4,5].

One of the principal selection mechanisms in different feature selection method is feature ranking. That is the features are ranked

* Corresponding author.
E-mail addresses: rasmitadash@soauniversity.ac.in, rasmita02@yahoo.co.in (R. Dash), misrabijan@gmail.com (Dr.B.B. Misra).

http://dx.doi.org/10.1016/j.asoc.2016.07.006
1568-4946/

based on some merit score computed based on ranking criterion and then few top ranked features are selected. It is a filter approach and has advantage over other feature selection scheme like it is simple to use, computationally and statistically scalable to large datasets and it has wide range of application.

Feature ranking method selects the significant features based on some selection criterion such as distance [6,7], information theory [8] or on some function of classifier's output. For example RELIEFF, which was proposed by Kira and Rendell [6] is one of the most successful distance based measure and adopt Euclidean distance to assign a relevance weight to each feature. But whatever may be the criterion, all conventional rank selection algorithms are based on single approach of evaluating the features. The difference in approach generates different ranks of features by different ranking methods. Therefore appropriate selection of features depends on the approach of evaluation of the ranking technique, which may favour the search space of the problem or may not. However, given a problem at hand, one does not possess the a priori knowledge about which criterion works the best for it. Therefore selection of appropriate ranking technique for the problem is an important as well as a difficult task.

This paper proposes pipelining of ranking techniques for efficient feature selection and classification of microarray databases. Four feature ranking methods are considered here for pipelining, such as information gain (IG), signal to noise ratio (SNR), Pearson correlation coefficient (PCC) and t-statistic. Based on the combination of 4 ranking methods, 24 pipelines are built up. Performance of the pipelines are evaluated using 4 classifiers, such as Multiple Linear Regression (MLR), Artificial Neural Network (ANN), Naïve Bayes network (NB) and k-Nearest Neighbour (k-NN). Further to find out which pipeline shows better performance, a grading method is used. In the first stage grading, the pipelines are graded with respect to different classifiers but result shows that the grading value varies from dataset to dataset. So it is difficult to conclude which pipeline performs the best. However, some of the pipelines perform better for most of the datasets, though not the best always. Therefore, grading in multiple stages is applied and the findings are validated using nonparametric statistical test.

Rest part of the paper is arranged as follows. In Section 2 recent literature studies on feature ranking method applied on microarray data is presented. In Section 3 the proposed model for feature selection and classification is discussed. The complete experimental work is described in Section 4. It includes the dataset used for the experimental analysis, data normalization process, dataset training, testing and validation process, feature ranking techniques used, performance matrices and result analysis. Finally the paper is concluded with Section 5.

2. Background study

The curse of dimensionality in microarray dataset makes it essential to have dimensionality reduction of the datasets before proceeding for classification. In this section different feature selection techniques are discussed with special reference to different ranking methods with their advantages and limitations.

2.1. Different feature selection techniques

Feature selection (or variable elimination) techniques are broadly categorised into two groups filter method and wrapper method [9]. The filter method considers the intrinsic properties of the data and generates a relevant set of features [10]. In this technique the genes are either ranked (depending on some scoring value) [11] or evaluated with respect to the cost function to identify the most informative genes [12]. The rank based filter method generates the relevant feature based on the intrinsic property of the dataset. In this technique features are evaluated on the basis of a scoring function where the top ranked features are selected and low ranked features are removed [11]. So to select the relevant feature a threshold value of the scoring function is chosen. The features with greater than or equal to the score value are selected and rest are removed. After that the subset features are presented for classification. The filter techniques are mostly effective as they are computationally fast and are easily scale to very high dimension data [13].

Table 1
List of ranking techniques applied to different microarray database.

Author	Feature Ranking Method	Reference
Hall et al.	Information Gain	[16]
Peyman et al.	*t*-test, ANOVA	[17]
M. A. Hall et al.	Correlation based feature selection (CBF)	[18]
Thomas et al., Tsai et al.	*t*-test	[19,20]
Thomas et al., Antoniadis et al.	Wilcoxon score test	[21,22]
Hwang et al	Wilks's Lambda score	[23]
Wang et al, Golub et al.	Signal to noise ratio	[24,25]
Cho et al, Ho et al.	Euclidian distance	[26,27]
Xing et al.	Information Gain	[28]

2.2. Feature selection using ranking methods in microarray databases

Recently, the ranked based feature selection has obtained more attention for solving the feature selection problem in many areas like sequence analysis, mass spectra analysis, single nucleotide polymorphisms (SNPs) analysis, text and literature mining, microarray data analysis and many more [12].

In many applications single ranking techniques applied to different microarray database listed in Table 1. Filter method rank each feature according to some univariate metric and only the highest ranking features are used while the remaining low ranking features are eliminated. This method also relies on general characteristics of the training data to select some features without involving any learning algorithm. Therefore, the results of filter model will not affect any classification algorithm. Moreover, filter methods also provide very easy way to calculate and can simply scale to large-scale microarray datasets since it has a short running time. In addition, filter methods also offer less computational time to generate results which is an extra point to be preferred by domain experts. However, gene ranking based on these techniques has some drawbacks. The major one is the genes selected are most probably redundant [14]. This is due to a rank method ranks the genes using single ranking criteria (or score function). But it is difficult to say which criteria suit a particular dataset to rank all the genes. Due to which some important genes may be rejected and some less important genes may be selected. To overcome this problem many embedded and ensemble gene selection techniques are applied to microarray databases [15].

Peng et al. [28] applied a hybrid approach using Fisher's ratio, a simple method easy to understand and implement, to filter out most of the irrelevant genes, then a wrapper method is employed to reduce the redundancy. The performance of FR-Wrapper approach is evaluated over four widely used microarray datasets (Leukemia, Lung cancer, Breast cancer and Colon cancer). Analysis of experimental results reveals that the hybrid approach can achieve the goal of maximum relevance with minimum redundancy.

A filter based on the *t*-statistic is used by Mundra et al. [29] in which t-statistic is divided into two parts, corresponding to *relevant* and *irrelevant* data points. The *relevant* data points are selected