Applied Soft Computing 17 (2014) 117–126

Contents lists available at ScienceDirect

Applied Soft Computing

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

A novel ensemble of classifiers that use biological relevant gene sets for microarray classification

Miguel Reboiro-Jato [a], Fernando Díaz [b], Daniel Glez-Peña [a], Florentino Fdez-Riverola [a,*]

[a] Departamento de Informática, Escuela Superior de Ingeniería Informática, University of Vigo, Ourense, Spain
[b] Departamento de Informática, Escuela Universitaria de Informática, University of Valladolid, Segovia, Spain

ARTICLE INFO

Article history:
Received 12 July 2011
Received in revised form 25 April 2013
Accepted 1 January 2014
Available online 17 January 2014

Keywords:
Microarray classification
Biological knowledge
Ensemble learning
Knowledge integration
Classifier fusion

ABSTRACT

Since the introduction of DNA microarray technology, there has been an increasing interest on clinical application for cancer diagnosis. However, in order to effectively translate the advances in the field of microarray-based classification into the clinic area, there are still some problems related with both model performance and biological interpretability of the results. In this paper, a novel ensemble model is proposed able to integrate prior knowledge in the form of gene sets into the whole microarray classification process. Each gene set is used as an informed feature selection subset to train several base classifiers in order to estimate their accuracy. This information is later used for selecting those classifiers comprising the final ensemble model. The internal architecture of the proposed ensemble allows the replacement of both base classifiers and the heuristics employed to carry out classifier fusion, thereby achieving a high level of flexibility and making it possible to configure and adapt the model to different contexts. Experimental results using different datasets and several gene sets show that the proposal is able to outperform classical alternatives by using existing prior knowledge adapted from publicly available databases.

1. Introduction

With the advent of microarray technology, tumor classification based on microarray gene expression data has become one of the most active research topics in bioinformatics. 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 drawbacks in achieving accurate and reproducible results. In order to obtain more robust microarray data classification methods, several authors have investigated the benefits of using ensemble alternatives applied to genomic research [2].

As a particular case in the use of ensemble systems, ensemble feature selection represents an efficient method proposed by Opitz [3] which can also achieve high classification accuracy by combining base classifiers built with different feature subsets. In this context, the works of Kuncheva and Jain [4] and Oliveira et al. [5] study the application of different genetic algorithms alternatives for performing feature selection with the aim of making classifiers of the ensemble disagree on difficult cases in order to introduce diversity. Reported results on both cases showed improvements when compared against other alternatives.

Díaz-Uriarte and Álvarez de Andrés [6] studied the use of random forests for multiclass classification of microarray data and proposed a new method of gene selection in classification problems based on this classifier. Using simulated and real microarray datasets, the authors showed that random forests can obtain comparable performance to other methods including DLDA (*Diagonal Linear Discriminant Analysis*), K-NN (*K-Nearest Neighbor*) and SVM (*Support Vector Machine*).

Peng [7] presented a novel ensemble approach based on seeking an optimal and robust combination of multiple classifiers. The proposed algorithm begins with the generation of a pool of candidate base classifiers based on gene sub-sampling and then, it performs the selection of a subset of appropriate base classifiers to construct the classification committee based on classifier clustering. Experimental results demonstrated that the proposed approach outperforms both baseline classifiers and classical ensemble algorithms, such as Bagging [8] and Boosting [9,10].

Liu and Huang [11] applied rotation forest to microarray data classification using principal component analysis, non-parametric discriminant analysis and random projections to perform feature transformation in the original rotation forest. In all the experiments, the authors reported that the proposed approach outperformed classical Bagging and Boosting alternatives.

* Corresponding author at: ESEI – Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004 Ourense, Spain. Tel.: +34 988 387015; fax: +34 988 387001.
E-mail addresses: mrjato@uvigo.es (M. Reboiro-Jato), fdiaz@infor.uva.es (F. Díaz), dgpena@uvigo.es (D. Glez-Peña), riverola@uvigo.es (F. Fdez-Riverola).

1568-4946/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2014.01.002

Chen and Zhao [12] presented an ensemble of classifiers based on correlation analysis for microarray data classification. Gene features are first extracted by correlation analysis in order to reduce dimensionality. Then, a pool of candidate base classifiers is generated to learn the subsets which are re-sampled with PSO (*Particle Swarm Optimization*). Finally, best classifiers are selected using EDAs (*Estimation of Distribution Algorithms*). They compared the results obtained against some advanced artificial techniques on four benchmark databases reporting best recognition rates.

More recent approaches include the works of Liu and Xu [13], Kodell et al. [14], Anand and Suganthan [15], Haferlach et al. [16], Zeng and Liu [17] and Yang et al. [18] that present different ensemble alternatives applied to microarray data classification by introducing variations at data, feature, classifier and combination levels [19].

Although numerical analysis of microarray data seems quite consolidated, the true integration of numerical analysis and biological knowledge is still a long way off [20]. In addition to classification performance, there is also hope that microarray studies uncover molecular disease mechanisms. However, in many cases the molecular signatures discovered by the algorithms are unfocused from a biological point of view [21]. In fact, they often look more like random gene lists than biologically plausible and understandable signatures. Moreover, an additional shortcoming of standard classification algorithms is that they treat gene-expression levels as anonymous attributes, but a lot is known about the function and the role of many genes in certain biological processes.

In this scenario, some authors have claimed that classifiers should take into account both model performance and biological interpretability of their results. With the goal of effectively incorporating prior knowledge into the classification process of microarray data, we have developed genEnsemble. The aim of the proposed model is to incorporate relevant gene sets obtained from our previous WhichGenes server [22] in order to make accurate predictions easy to interpret in concert with existing knowledge. The study carried out in this research work was inspired in past successful results [23] and aims to borrow information from existing biological knowledge to improve both predictive accuracy and interpretability of the final generated classifier.

The structure of the paper is as follows. Section 2 presents previous approaches that integrate biological knowledge in the classification process by using ensemble alternatives. Section 3 describes the proposed knowledge-based ensemble model and details how base classifiers are initially selected and further combined. Section 4 introduces the experimental framework, discusses the obtained results and analyses the main characteristics of the generated ensemble. Finally, Section 5 outlines main conclusions and identifies future research lines.

2. Integration of knowledge into the classification process

As previously stated, the inclusion of additional knowledge sources in the microarray classification process can alleviate several problems related with model performance (e.g.: overfitting and/or the inability of the model to generalize properly) while improving interpretability of results. Related with this issue, the integration of knowledge can prevent the discovery of the obvious, complement a data-inferred hypothesis with references to already proposed relations, avoid overconfident predictions and allow to systematically relate the analysis findings to present knowledge [24]. Mainly motivated by the previous commented advantages, some authors have proposed different alternatives for the integration and use of external knowledge in ensemble-based predictors.

In this context, the work of Pang et al. [25] combines a random forests classifier with pathway information. The main objective of this work is to explore the capacity of random forests to assess the importance of variables in order to derive the enrichment of pathways. In this sense, the work is more focused on the functional analysis than in the classification task.

Later, Wei and Li developed NPR (*Nonparametric Pathway-based Regression*) [26], a modification of the Boosting schema. This work proposes that, at each step of the boosting procedure, a classifier focused on each metabolic pathway is trained and the best one is selected. At the end of the process, a classifier based on several base learners with biological criteria is obtained and the best pathways during the boosting are also reported. During their experiments, the authors included different pathways related to breast cancer disease.

More recently, Binder and Schumacher developed PathBoost [27], an improved version of their BRR (*Boosting Ridge Regression*) algorithm previously proposed [28]. BRR is based on Boosting, but also incorporates a gene selection schema. PathBoost improves BRR by incorporating gene networks in order to jointly select those network-related genes. They derived a gene network from KEGG by creating a gene–gene connection every time there are relations in the KEGG database.

Finally, Daemen et al. [29] proposed an ensemble model able to integrate three different data sources (metabolic pathway information, protein-protein interactions and miRNA-gene targeting) by exploiting the properties of kernel methods. They incorporated the relations between genes with similar functions but active in alternative pathways in a LS-SVM (*Least Squares SVM*) classifier using spectral graph theory. The graph-related information was subsequently utilized by the SVM for the calculation of patient similarity.

Although all the previous approaches reported valuable improvements related with model performance or interpretability, they are constrained by the use of a few limited sources of biological information or the utilization of a specific classifier/algorithm.

3. genEnsemble: using prior biological knowledge for microarray classification

Taking into consideration the state-of-the-art and with the goal of overcoming existing limitations of previous works we have developed genEnsemble, a classification model in which the knowledge is freely represented using biologically relevant and problem-related gene sets. This set-based approach allows both (i) the homogenization of knowledge coming from multiple and different sources of biological information in an intuitive way for the expert and transparent for the model and (ii) the effectively use of this knowledge as a natural way of performing the necessary feature selection process for microarray classification.

By following this straightforward approach, the proposed model is able to integrate data and knowledge through the training of several classifiers that use the initial genet sets provided by the expert as a feature selection filter directly applied to the input microarray samples. With the goal of guaranteeing the diversity in the generated ensemble, the proposed model is able to use different types of base classifiers due to the fact that it defines a heterogeneous model for the combination of classification techniques based on abstract outputs (i.e.: taking into consideration only the class label assigned by each base classifier).

As a result, given a microarray dataset and a group of gene sets provided by the expert, there are four main steps required by genEnsemble in order to construct the final classification model: (i) generation of candidate classifiers, (ii) evaluation of candidate classifiers, (iii) selection of base classifiers and (iv) training of base classifiers. The following subsections introduce and explain each phase in detail.