Expert Systems with Applications 40 (2013) 52–63

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

genEnsemble: A new model for the combination of classifiers and integration of biological knowledge applied to genomic data

Miguel Reboiro-Jato [a], Rosalía Laza [a], Hugo López-Fernández [a], Daniel Glez-Peña [a], Fernando Díaz [b], Florentino Fdez-Riverola [a,*]

[a] Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, University of Vigo, 32004 Ourense, Spain
[b] Escuela Universitaria de Informática, Plaza de Santa Eulalia 9-11, University of Valladolid, 40005 Segovia, Spain

ARTICLE INFO

Keywords:
Ensemble approaches
Microarray data classification
Knowledge integration
Inter-dataset robustness

ABSTRACT

In the last years, microarray technology has become widely used in relevant biomedical areas such as drug target identification, pharmacogenomics or clinical research. However, the necessary prerequisites for the development of valuable translational microarray-based diagnostic tools are (i) a solid understanding of the relative strengths and weaknesses of underlying classification methods and (ii) a biologically plausible and understandable behaviour of such models from a biological point of view. In this paper we propose a novel classifier able to combine the advantages of ensemble approaches with the benefits obtained from the true integration of biological knowledge in the classification process of different microarray samples. The aim of the current work is to guarantee the robustness of the proposed classification model when applied to several microarray data in an inter-dataset scenario. The comparative experimental results demonstrated that our proposal working with biological knowledge outperforms other well-known simple classifiers and ensemble alternatives in binary and multiclass cancer prediction problems using publicly available data.

1. Introduction

Nowadays, the popularization and use of DNA microarray technology in clinics, medical centres and hospitals makes possible the measurement of gene expression profiles and their translational application for clinical decision support (Quackenbush, 2006). In this context, and with the goal of correctly identifying and diagnosing different complex diseases – such as cancer – a large group of computational methods ranging from statistical alternatives to more complex hybrid systems have been previously proposed in the literature (Ressom, Varghese, Zhang, Xuan, & Clarke, 2008).

From all the available alternatives, and taking into consideration the inherent high dimensional input space problem of microarray data, two-stage machine learning classification techniques are particularly suitable for producing accurate disease classifications as well as competent prognosis outcomes (Wong & Hsu, 2008). Following this schema, a first gene selection stage is carried out with the goal of both reducing data dimensionality and improving classifier performance by avoiding the over-fitting problem (Horng et al., 2009). Apart from the specific use of an appropriate gene selection mechanism – based on subset or individual gene ranking (Zhang, Lu, & Zhang, 2006) – an adequate classification algorithm is needed to produce a high prediction accuracy of classification outcomes (Pirooznia, Yang, Yang, & Deng, 2008).

In this context, both well-known simple approaches and more accurate ensemble of classifiers have been previously applied to microarray classification with different levels of success (Statnikov, Aliferis, Tsamardinos, Hardin, & Levy, 2005). In the first group, remarkable stand-alone techniques are support vector machines (SVM) (Nanni & Lumini, 2010), back-propagation neural networks (BPNN) (Chen & Lin, 2011), naïve Bayes (NB) (Fan, Poh, & Zhou, 2010), wavelet neural networks (WNN) (Zainuddin & Ong, 2011), random forests (RF) (Statnikov, Wang, & Aliferis, 2008), decision trees (DT) (Che, Liu, Rasheed, & Tao, 2011) or the simple k-NN algorithm (Parry et al., 2010). In the second group, ensembles of classifiers have gained increased importance because they have shown to perform better than stand-alone alternatives (Chen et al., 2011; Moon et al., 2007). The main idea behind this approach is that by averaging the different output of base classifiers, the combined systems will produce a good approximation of the true hypothesis (Nanni & Lumini, 2009). From all the possibilities, there are three basic alternatives that stand out for effectively integrating individual classifier decisions (Kuncheva, 2004): Bagging (Breiman, 1996), Boosting (Schapire, 1990) and Stacking (Wolpert, 1992).

* Corresponding author. Address: 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), rlaza@uvigo.es (R. Laza), hlfernandez@sing.ei.uvigo.es (H. López-Fernández), dgpena@uvigo.es (D. Glez-Peña), fdiaz@infor.uva.es (F. Díaz), riverola@uvigo.es (F. Fdez-Riverola).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.07.003

However, as previously stated by Lottaz and Spang (2005), in addition to the predictive performance of available models there is also hope that microarray classification studies uncover the underlying molecular disease mechanisms allowing biologically resolved diagnosis of patients. Nevertheless, extracting clear and coherent biological hypotheses from microarray classification experiments still remains an important challenge (Cordero, Botta, & Calogero, 2007). This situation is mainly motivated because (i) commonly applied gene selection techniques usually treat gene-expression levels as anonymous variables – although a lot is known about the function and role of many genes in certain biological processes – and (ii) existing microarray classification models are usually trained and tested following an intra-dataset schema – leading to overconfident results not generalizable to any new datasets -. Moreover, experience also demonstrates that classification performance in a microarray experiment strongly depends on the dataset, and less on the variable selection and classification method (Rocke, Ideker, Troyanskaya, Quackenbush, & Dopazo, 2009).

In such a situation, it is widely recognized that the integration of additional sources of biological knowledge in the classification process can (i) prevent the discovery of the obvious, (ii) complement a data-inferred hypothesis with references to already proposed relations, (iii) help analysis to avoid overconfident predictions and (iv) allow to systematically relating the analysis findings to present knowledge (Bellazzi & Zupan, 2007). Although the utilization of biological knowledge is usually associated with post-analyses studies using bioinformatics tools, the use of biology-driven filters as prior gene selection mechanism could be useful when biologists and clinicians have a clear idea of which subareas of biology are related with the different clinical phenotypes under study. These types of filters are clearly dependent on the availability of biological sources of knowledge and its relevance to the accessible microarray data, but they make possible to strengthen the biological mean of obtained results by increasing the robustness of the selected models when they are applied over different datasets.

With the goal of aiding in the generation of biologically resolved classifications using microarray data without compromising overall prediction accuracy of machine learning techniques we propose 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.

The paper is organized as follows: in Section 2 we summarize previous work integrating biological knowledge for gene selection, clustering and classification using microarray data. In section 3 the proposed approach is presented in detail, covering the methods and knowledge used in its definition. Section 4 introduces the experimental framework and discusses the results obtained from different points of view. Finally, in Section 5 some concluding remarks are given and future work is outlined.

2. Previous studies on knowledge integration for microarray data

Some previous works have successfully studied the integration of prior knowledge in different areas of microarray analysis including (i) gene selection (or biomarker discovery), (ii) co-expression analysis (or gene clustering) and (iii) sample classification.

On the one hand, the gene selection task focuses at discovering genes whose expression level is consistently different between phenotypes or sample conditions, also known as differential expressed genes. Classical approaches use statistical tests, such as *t*-test or ANOVA F, plus a multiple test correction step in order to find such biomarkers. In this context, Morrison, Breitling, Higham, and Gilbert (2005) and Daigle and Altman (2008) have proposed the inclusion of gene networks as previous knowledge. Based on both available protein-protein interaction databases and/or synthesised knowledge from gene annotations – such as metabolic pathway databases –, these approaches try to discover genes with models statistical power but being tightly connected to differential expressed genes in the knowledge networks.

On the other hand, the co-expression analysis aims at finding groups of genes with similar expression pattern among samples. This task is usually carried out by using clustering techniques (e.g., hierarchical algorithms) where Euclidean distance and Pearson correlation coefficient are common measures used to build clusters. In addition to purely data-based measures, Kustra and Zagdański (2007) have proposed a distance function based on the functional similarity of genes via Gene Ontology (GO) (The Gene Ontology Consortium., 2000). Similar to co-expression analysis, biological network inference from microarray data has been successfully enhanced by using prior knowledge in the form of graphs and Bayesian networks in the works of Imoto et al. (2003) and Tamada et al. (2003).

Finally, on the phenotype or condition prediction task, which is the main topic of the present work, models are built on top of microarray training data in order to predict the class of unseen or testing samples. Although classical techniques have been extensively applied to microarray classification tasks, (i) the overestimated performance of these studies (Wood, Visscher, & Kerrie, 2007), (ii) the overfitting problem derived from the high noise level in the form of dozens of thousands of non-related input genes compared to the low number of samples and (iii) the lack of biological interpretation of the selected genes as biomarkers (Quackenbush, 2006) have become the main difficulties in order to effectively translate microarray technology as a confident diagnostic tool to the clinic area.

Nevertheless, in the line of our proposed model, knowledge integration has also been introduced to build more robust and interpretable microarray classifiers. In this context, relevant improvements of the penalty-based PAM and PLS methods were proposed by Tai and Pan (2007a), where the penalty value is less stringent for functionally known genes. The successful incorporation of metabolic pathways in order to improve a simple LDA classifier was proposed by Tai and Pan (2007b) and the enhancement of SVMs was also studied by Zhu, Shen, and Pan (2009) and Guan, Huang, He, and Zhou (2009). More recently, Papachristoudis, Diplarisb, and Mitkasb (2010) used the GO ontology in their SoFoCles system to enrich the feature selection step with functional similar genes before forwarding them to a common classifier.

3. Proposed approach: methods and knowledge

As stated before, in this work we present genEnsemble, a novel and flexible model able to effectively integrate prior knowledge into the classification process of microarray data with the goal of both (i) guaranteeing the robustness of the generated classifications when applied to inter-dataset scenarios and (ii) enhancing the biological interpretation of the results achieved by the system.

In genEnsemble, prior knowledge initially provided by the expert is structured in the form of input gene sets, which are associated with user-dependant concepts of interest (e.g.: diseases, pathways, etc.) related to the microarray samples under study. Both, structured knowledge (gene sets) and microarray samples (dataset) are required inputs of the proposed system (see Fig. 1).