A Comparative Study of Microarray Data Classification Methods Based on Ensemble Biological Relevant Gene Sets

Miguel Reboiro-Jato, Daniel Glez-Peña, Juan Francisco Gálvez,
Rosalía Laza Fidalgo, Fernando Díaz, and Florentino Fdez-Riverola

Abstract. In this work we study the utilization of several ensemble alternatives for the task of classifying microarray data by using prior knowledge known to be biologically relevant to the target disease. The purpose of the work is to obtain an accurate ensemble classification model able to outperform baseline classifiers by introducing diversity in the form of different gene sets. The proposed model takes advantage of WhichGenes, a powerful gene set building tool that allows the automatic extraction of lists of genes from multiple sparse data sources. Preliminary results using different datasets and several gene sets show that the proposal is able to outperform basic classifiers by using existing prior knowledge.

Keywords: microarray data classification, ensemble classifiers, gene sets, prior knowledge.

1 Introduction and Motivation

The advent of microarray technology has become a fundamental tool in genomic research, making it possible to investigate global gene expression in all aspects of human disease. In particular, cancer genetics based on the analysis of cancer genotypes, provides a valuable alternative to cancer diagnosis in both theory and practice [1]. In this context, the automatic classification of cancer patients has been a

Miguel Reboiro-Jato · Daniel Glez-Peña · Juan Francisco Gálvez · Rosalía Laza Fidalgo · Florentino Fdez-Riverola
ESEI: Escuela Superior de Ingeniería Informática, University of Vigo,
Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004, Ourense, Spain
e-mail: {mrjato, dgpena, galvez, rlaza, riverola}@uvigo.es

Fernando Díaz
EUI: Escuela Universitaria de Informática, University of Valladolid, Plaza Santa Eulalia, 9-11, 40005, Segovia, Spain
e-mail: fdiaz@infor.uva.es

M.P. Rocha et al. (Eds.): IWPACBB 2010, AISC 74, pp. 25–32, 2010.
springerlink.com

promising approach in cancer diagnosis since the early detection and treatment can substantially improve the survival rates. For this task, several computational methods (statistical and machine learning) have been proposed in the literature including linear discriminant analysis (LDA), Naïve-Bayes classifier (NBC), learning vector quantization (LVQ), radial basis function (RBF) networks, decision trees, probabilistic neural networks (PNNs) and support vector machines (SVMs) among others [2]. In the same line, but following the assumption that a classifier ensemble system is more robust than an excellent single classifier [3], some researchers have also successfully applied different classifier ensemble systems to deal with the classification of microarray datasets [4].

In addition to predictive 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 form a biological point of view [5]. In fact, they often look more like random gene lists than biologically plausible and understandable signatures. Another shortcoming of standard classification algorithms is that they treat gene-expression levels as anonymous attributes. However, a lot is known about the function and the role of many genes in certain biological processes.

Although numerical analysis of microarray data is considerable consolidated, the true integration of numerical analysis and biological knowledge is still a long way off [6]. The inclusion of additional knowledge sources in the classification process can prevent the discovery of the obvious, complement a data-inferred hypothesis with references to already proposed relations, help analysis to avoid overconfident predictions and allow us to systematically relate the analysis findings to present knowledge [7]. In this work we would like to incorporate relevant gene sets obtained from WhichGenes [8] in order to make predictions easy to interpret in concert with incorporated knowledge. The study carried out aims to borrow information from existing biological knowledge to improve both predictive accuracy and interpretability of the resulting classifiers.

The rest of the paper is structured as follows: Section 2 presents a brief review about the use of ensemble methods for classifying microarray data. Section 3 describes the selected datasets and base classifiers for the current study, together with the choice of gene sets and the different approaches used for ensemble creation. Finally Section 4 discusses the reported results and concludes the paper.

2 Related Work

Although much research has been performed on applying machine learning techniques for microarray data classification during the past years, it has been shown that conventional machine learning techniques have intrinsic drawbacks in achieving accurate and robust classifications. In order to obtain more robust microarray data classification techniques, several authors have investigated the benefits of this approach applied to genomic research.

Díaz-Uriarte and Alvarez de Andrés [9] investigated the use of random forest for multi-class classification of microarray data and proposed a new method of gene selection in classification problems based on random forest. Using simulated

and real microarray datasets the authors showed that random forest can obtain comparable performance to other methods, including DLDA, KNN, and SVM.

Peng [10] 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 the gene subsampling and then, it performs the selection of a sub-set 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 those generated by bagging and boosting.

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 bagging and boosting alternatives.

More recently, Liu and Xu [12] proposed a genetic programming approach to analyze multiclass microarray datasets where each individual consists of a set of small-scale ensembles containing several trees. In order to guarantee high diversity in the individuals a greedy algorithm is applied. Their proposal was tested using five datasets showing that the proposed method effectively implements the feature selection and classification tasks.

As a particular case in the use of ensemble systems, ensemble feature selection represents an efficient method proposed in [13] which can also achieve high classification accuracy by combining base classifiers built with different feature subsets. In this context, the works of [14] and [15] study the use of different genetic algorithms alternatives for performing feature selection with the aim of making classifiers of the ensemble disagree on difficult cases. Reported results on both cases showed improvements when compared against other alternatives.

Related with previous work, the aim of this study is to validate the superiority of different classifier ensemble approaches when using prior knowledge in the form of biological relevant gene sets. The objective is to improve the predictive performance of baseline classifiers.

3 Comparative Study

In order to carry out the comparative study, we apply several ensemble alternatives to classify three DNA microarray datasets involving various tumour tissue samples. With the goal of validate the study, we analyze the performance of different baseline classifiers and test our hypothesis using two different sources of information.

3.1 Datasets and Base Classifiers

We carry out the experimentation using three public leukemia datasets taken from the previous studies of Gutiérrez *et al* [16], Bullinger *et al* [17] and Valk *et al* [18]. We have selected samples from each dataset belonging to 4 different groups of acute myeloid leukemias including (*i*) promyelocytic (APL), (*ii*) inversion 16, (*iii*) monocytic and (*iv*) other AMLs. The distribution of samples is showed in Table 1.

Table 1 Distribution of microarray data samples belonging to the public datasets analyzed

	APL	Inv(16)	Monocytic	Other
Gutiérrez *et al*	10	4	7	22
Bullinger *et al*	19	14	64	177
Valk *et al*	7	10	7	51

In order to compare the performance obtained by the different ensemble approaches, we have selected four well-known classification algorithms: (*i*) Naïve Bayes (NB) learner is perhaps the most widely used method. Although its independence assumption is over-simplistic, studies have found NB to be very effective in a wide range of problems; (*ii*) IB3 represents a variant of the well-known nearest neighbour algorithms implementing a simple version of a lazy learner classifier; (*iii*) Support Vector Machines (SVMs) constitute a famous family of algorithms used for classification and regression purposes. Their mayor advantage is that their learning capacity does not degrade even if many characteristics exist, being especially applicable to microarray data; (*iv*) Random Forest (RFs) is a basic ensemble classifier that consists of many decision trees. The method combines bagging idea and random selection of features in order to construct a collection of decision trees with controlled variation.

3.2 Biological Knowledge Gene Sets

For the prior selection of gene sets that represent explicit information available the following sources of information have been used: (*i*) 33 metabolic sub-pathways related to existing cancers in SABiosciences (http://www.sabiosciences.com) previously analyzed in studies by [19] and [20] plus 4 groups extracted from the OMIM (*Online Mendelian Inheritance in Man*) database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim) that correspond to various types of leukemia (myeloid, monocytoid, promyelocytic and general leukemia) and (*ii*) those pathways from KEGG (*Kyoto Encyclopedia of Genes and Genomes*) database grouped in both 'environmental information processing' and 'genetic information processing' categories.

3.3 Ensemble Alternatives

According to Kuncheva [3], several ensembles can be built by introducing variations at four different levels: (*i*) data level, (*ii*) feature level, (*iii*) classifier level and (*iv*) combination level.

First of all, by using different data subsets at data level or different feature subsets at feature level, the space of the problem can be divided into several areas where base classifiers can be trained. This divide-and-conquer strategy can simplify the problem, leading to improved performance of the base classifiers. Secondly, at classifier level, different types of classifiers can be used in order to take advantage of the strong points of each classifier type. Although many ensemble paradigms employ the same classification model, there is no evidence that one strategy is better than the other [3]. Finally, combination level groups the different ways of combining the classifier decisions.

In this study, base classifiers are trained with all the samples in each data set, so no work is performed at data level. The feature level is carried out by incorporating gene set data to the ensemble models. Each pathway or group of genes is used as a feature selection, so microarray data will be filtered to keep only the expression level of those genes belonging to some group before training base classifiers.

In order to construct the final ensemble, our approach consists on two sequential steps: (*i*) *classifier selection*, in which each simple classifier is initially trained with each gene set following a stratified 10-fold cross-validation process for estimating its performance and (*ii*) *classifier training*, where the selected pairs of simple_classifier/gene_set are trained with the whole data set. All the different strategies proposed in this study for the selection of promising classifiers are based on the value of the kappa statistic obtained for each simple_classifier/gene_set pair in the first step. The proposed heuristics are the following:

- *All classifiers* [AC]: every simple_classifier/gene_set pair is used for constructing the final ensemble.
- *All gene sets* [AG]: for each gene set, the simple_classifier/gene_set pair with best kappa value is selected for constructing the final ensemble.
- *Best classifiers without type* [BCw/oT_%]: a global threshold is calculated as a percentage of the best kappa value obtained by the winner simple_classifier/gene_set pair. Those pairs with a kappa value equal or higher than the computed threshold are selected.
- *Best classifier by type* [BCbyT_%]: as in the previous heuristic a given threshold is calculated, but in this case there is a threshold for each simple classifier type.

The form in which the final output of the ensemble is calculated is also based on the kappa statistic. The combination approach used on for the proposed ensembles is a weighted majority vote where the weight of each vote is the corresponding classifier's kappa value.

4 Experimental Results and Discussion

In order to evaluate the heuristics defined in the previous section, a comparative study was carried out using two different sources of information (OMIM and KEGG) in order to classify 392 samples belonging to four classes coming from three real data sets. In addition, the four simple base classifiers used for the ensemble generation (IB3, NBS, RF, SVM) where also tested individually, using as feature selection both those genes included in the OMIM gene sets plus those genes being part of the KEGG gene sets. Classification tests were performed using a stratified 10-fold cross-validation. Tables 2 and 3 summarize the results obtained from the experimentation carried out showing only those classifiers with better performance.

Table 2 presents the accuracy and kappa values achieved by each classifier using KEGG gene sets as prior knowledge. As it can be observed, BCbyT heuristic generally exhibits good performance regardless of the data set. Additionally, BCw/oT heuristic also showed good performance, although in the Gutiérrez data set two single classifiers (IB3 and NBS) performed better than ensembles using this strategy.

Table 2 Classification result using KEGG gene sets

Classifier	Gutiérrez		Bullinger		Valk	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
AC	76,74%	0,588	76,00%	0,292	76,28%	0,503
AG	79,07%	0,634	76,00%	0,299	75,18%	0,533
BCbyT_90%	81,40%	0,724	82,67%	0,373	77,01%	0,540
BCbyT_95%	83,72%	0,724	82,67%	0,329	**78,83%**	**0,574**
BCw/oT_60%	79,07%	0,635	**80,00%**	**0,476**	77,37%	0,556
BCw/oT_75%	79,07%	0,635	81,33%	0,367	76,64%	0,555
BCw/oT_85%	76,74%	0,612	80,00%	0,403	75,55%	0,546
IB3	**83,72%**	**0,756**	69,33%	0,369	67,52%	0,410
NBS	81,40%	0,679	73,33%	0,269	74,09%	0,530
RF	72,09%	0,533	68,00%	0,123	69,71%	0,337
SVM	51,16%	0,000	68,00%	0,000	64,60%	0,000

Table 3 Classification result using OMIM gene sets

Classifier	Gutiérrez		Bullinger		Valk	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
AC	76,74%	0,588	76,00%	0,343	74,82%	0,439
AG	76,74%	0,588	76,00%	0,343	76,28%	0,506
BCbyT_90%	81,40%	0,680	**82,67%**	**0,569**	75,91%	0,528
BCbyT_95%	86,05%	0,774	**82,67%**	**0,569**	75,91%	0,513
BCw/oT_60%	81,40%	0,680	80,00%	0,483	75,91%	0,521
BCw/oT_75%	**88,37%**	**0,809**	81,33%	0,526	75,91%	0,530
BCw/oT_85%	79,07%	0,672	80,00%	0,482	**76,28%**	**0,555**
IB3	76,74%	0,643	73,33%	0,451	67,52%	0,391
NBS	79,07%	0,634	76,00%	0,370	72,99%	0,510
RF	79,07%	0,658	74,67%	0,372	74,09%	0,420
SVM	51,16%	0,000	68,00%	0,000	77,74%	0,539

Table 3 presents the same experimentation but using the OMIM gene sets. Once again, BCbyT heuristic achieved good performance. Comparing its behaviour against single classifiers, performance of ensembles is even better than in the previous experimentation (using KEGG gene sets). BCw/oT heuristic also performs better with the OMIM gene set, being slightly superior to BCbyT heuristic. Ensembles using this strategy not only performed better than single classifiers, but also achieved the best kappa value in two of the three analyzed data sets.

To sum up, we can conclude that BCbyT heuristic performed as the best base classifier selection strategy, followed closely by *BCw/oT* heuristic. This fact backs up the following ideas: (*i*) depending on the data set there is not a single classifier able to achieve good performance in concert with the supplied knowledge and (*ii*) the presence of each classifier type in the final ensemble may improve the classification performance.

Regardless of the data set both *BCw/oT* and BCbyT heuristics behave uniformly performing better than single baseline classifiers. This circumstance

confirms the fact that ensembles generally perform better than single classifiers, in this case, by taking advantage of using prior structured knowledge.

Acknowledgements. This work is supported in part by the project *MEDICAL-BENCH: Platform for the development and integration of knowledge-based data mining techniques and their application to the clinical domain* (TIN2009-14057-C03-02) from Ministerio de Ciencia e Innovación (Spain). D. Glez-Peña acknowledges Xunta de Galicia (Spain) for the program Ángeles Álvariño.

References

1. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
2. Ressom, H.W., Varghese, R.S., Zhang, Z., Xuan, J., Clarke, R.: Classification algorithms for phenotype prediction in genomics and proteomics. Frontiers in Bioscience 13, 691–708 (2008)
3. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley Interscience, Hoboken (2004)
4. Liu, K.H., Li, B., Wu, Q.Q., Zhang, J., Du, J.X., Liu, G.Y.: Microarray data classification based on ensemble independent component selection. Computers in Biology and Medicine 39(11), 953–960 (2009)
5. Lottaz, C., Spang, R.: Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data. Bioinformatics 21(9), 1971–1978 (2005)
6. Cordero, F., Botta, M., Calogero, R.A.: Microarray data analysis and mining approaches. Briefings in Functional Genomics and Proteomics 6(4), 265–281 (2007)
7. Bellazzi, R., Zupan, B.: Methodological Review: Towards knowledge-based gene expression data mining. Journal of Biomedical Informatics 40(6), 787–802 (2007)
8. Glez-Peña, D., Gómez-López, G., Pisano, D.G., Fdez-Riverola, F.: WhichGenes: a web-based tool for gathering, building, storing and exporting gene sets with application in gene set enrichment analysis. Nucleic Acids Research 37(Web Server issue), W329–W334 (2009)
9. Díaz-Uriarte, R., Alvarez de Andrés, S.: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3 (2006)
10. Peng, Y.: A novel ensemble machine learning for robust microarray data classification. Computers in Biology and Medicine 36(6), 553–573 (2006)
11. Liu, K.H., Huang, D.S.: Cancer classification using Rotation Forest. Computers in Biology and Medicine 38(5), 601–610 (2008)
12. Liu, K.H., Xu, C.G.: A genetic programming-based approach to the classification of multiclass microarray datasets. Bioinformatics 25(3), 331–337 (2009)
13. Opitz, D.: Feature selection for ensembles. In: Proceedings of 16th National Conference on Artificial Intelligence, Orlando, Florida (1999)
14. Kuncheva, L.I., Jain, L.C.: Designing classifier fusion systems by genetic algorithms. IEEE Transactions on Evolutionary Computation 4(4), 327–336 (2000)

15. Oliveira, L.S., Morita, M., Sabourin, R.: Feature selection for ensembles using the multi-objective optimization approach. Studies in Computational Intelligence 16, 49–74 (2006)
16. Gutiérrez, N.C., López-Pérez, R., Hernández, J.M., Isidro, I., González, B., Delgado, M., Fermiñán, E., García, J.L., Vázquez, L., González, M., San Miguel, J.F.: Gene expression profile reveals deregulation of genes with relevant functionsin the different subclasses of acute myeloid leukemia. Leukemia 19(3), 402–409 (2005)
17. Bullinger, L., Döhner, K., Bair, E., Fröhling, S., Schlenk, R.F., Tibshirani, R., Döhner, H., Pollack, J.R.: Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. The New England Journal of Medicine 350(16), 1506–1516 (2004)
18. Valk, P.J., Verhaak, R.G., Beijen, M.A., Erpelinck, C.A., Barjesteh van Waalwijk van Doorn-Khosrovani, S., Boer, J., Beverloo, H., Moorhouse, M., van der Spek, P., Löwenberg, B., Delwel, R.: Prognostically useful gene-expression profiles in Acute Myeloid Leukemia. The New England Journal of Medicine 350(16), 1617–1628 (2004)
19. Tai, F., Pan, W.: Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. Bioinformatics 23(14), 1775–1782 (2007)
20. Wei, Z., Li, H.: Nonparametric pathway-based regression models for analysis of genomic data. Biostatistics 8(2), 265–284 (2007)