A gene expression analysis system for medical diagnosis

Dimitris Maroulis[1], Dimitris Iakovidis[1], Stavros Karkanis[2], Ilias Flaounas[1]
1 University of Athens, Dept. of Informatics and Telecommunications, Panepistimiopolis, 15784 Ilisia, Greece
2 Lamia Institute of Technology, Dept. of Informatics and Computer Technology, 35100 Lamia, Greece
rtsimage@di.uoa.gr

Abstract. In this paper we present a novel system that utilizes molecular-level information for medical diagnosis. It accepts high dimensional vectors of gene expressions, quantified by means of microarray image analysis, as input. The proposed system incorporates various data pre-processing methods, such as missing values estimation and data normalization. A novel approach to the classification of gene expression vectors in multiple classes that embodies vari-ous gene selection methods has been adopted for diagnostic purposes. The pro-posed system has been extensively tested on various, publicly available data-sets. We demonstrate its performance for prostate cancer diagnosis and com-pare its performance with a well established multiclass classification scheme. The results show that the proposed system could be proved a valuable diagnostic aid in medicine.

1 Introduction

Microarray analysis has yet to be widely accepted for diagnosis and classification of diseases, despite the exponential increase in microarray studies reported in the literature. In the last decade a variety of software systems dedicated to microarray analysis have been developed. Do et al. [1] proposed the GeneClust software for microarray data analysis which implements hierarchical clustering and gene shaving algorithms [2]. Li and Wong [3] proposed the dChip software which implements a model-based expression analysis of oligonucleotide arrays and several high-level analysis procedures, including comparative analysis and hierarchical clustering. Peterson [4] proposed Clusfavor, a software package oriented in unsupervised analysis of microarrays. A powerful software suite named Genesis has been developed by Sturn et al. [5] for large-scale gene expression analysis. It includes filters, normalization and visualization tools, distance measures as well as clustering and classification algorithms such as hierarchical clustering, self-organizing maps, k-

Please use the following format when citing this chapter:
Maroulis, Dimitris, Iakovidis, Dimitris, Karkanis, Stavros, Flaounas, Ilias, 2006, in IFIP International Federation for Information Processing, Volume 204, Artificial Intelligence Applications and Innovations, eds. Maglogiannis, I., Karpouzis, K., Bramer, M., (Boston: Springer), pp. 459–466

means, principal component analysis, and Support Vector Machines (SVMs). Colantuoni et al. [6] developed a web-based tool named Snomad for the standardization and normalization of microarray data, using two non-linear transformations which correct both bias and variance of microarray element signal intensities. Saal et al. [7] developed Base, a software system for the management of biomaterial information, raw data and images, which provides integrated and "plug-in"-able normalization, data viewing and analysis tools. An open source suite of tools named TM4 has been developed by Saeed et al. [8]. It consists of four major applications, namely Microarray Data Manager (MADAM) which is a data entry and management tool for microarray experiments, TIGR_Spotfinder which is a semi-automated image analysis software, Microarray Data Analysis System (MIDAS) which is used for data normalization and filtering, and Multiexperiment Viewer (MeV) which is a data mining tool that implements a variety of clustering algorithms. Another software suite provided in open source is Bioconductor. It comprises of several packages that provide innovative tools for the analysis and comprehension of genomic data [9]. Su et al. developed RankGene, a software system which integrates a variety of popular ranking criteria, ranging from the traditional *t*-statistic to the one-dimensional SVMs [10]. A minimum spanning tree representation of gene expression data is being exploited by Excavator, a software system for microarray data clustering [11]. Toyoda and Konagaya [12] developed KnowledgeEditor, a graphical aid for biologists on biomolecular network modelling. Recently Pieler et al. proposed ArrayNorm a versatile and platform-independent application for the visualization, normalization and statistical identification of genes with significant changes in expression [13].

Most of the available software systems require technical skills and knowledge of complicated operations with which, physicians and biologists are not usually familiar. In this paper we present a novel, user friendly microarray data analysis software system which utilizes gene expression data for medical diagnosis. The proposed system does not require any technical knowledge by its users. It implements various pre-processing methods, and features a novel SVM-based architecture that embodies various gene selection methods in its structure and allows for the discrimination of multiple diseases or subtypes of a disease. Moreover, it handles the adjustment of its parameters automatically.

The rest of this paper is organized in four sections. Section 2 provides an overview of the proposed system and describes the methods it embodies. Experimental results from its application for prostate cancer diagnosis are presented in Section 3. Finally, in the last section the conclusions of this study are summarized.

2 System's Overview

The proposed system is capable of "learning" to recognize the pathology of samples provided to its input through a supervised training procedure. It embodies a Pre-processing and a Diagnostic Unit. The Pre-processing Unit prepares the gene expression data to passing into the Diagnostic Unit, which is the main processing unit of the proposed system.

The system's graphical user interface (Fig. 1) allows the user to switch between two modes of operation: the training and the diagnostic mode. The training mode of operation requires a gene expression matrix of pathologically characterized samples as input. The system automatically determines the best parameter settings for a particular diagnostic problem by grid search. After training, the system is capable of performing medical diagnosis based on a patient's gene expression data.

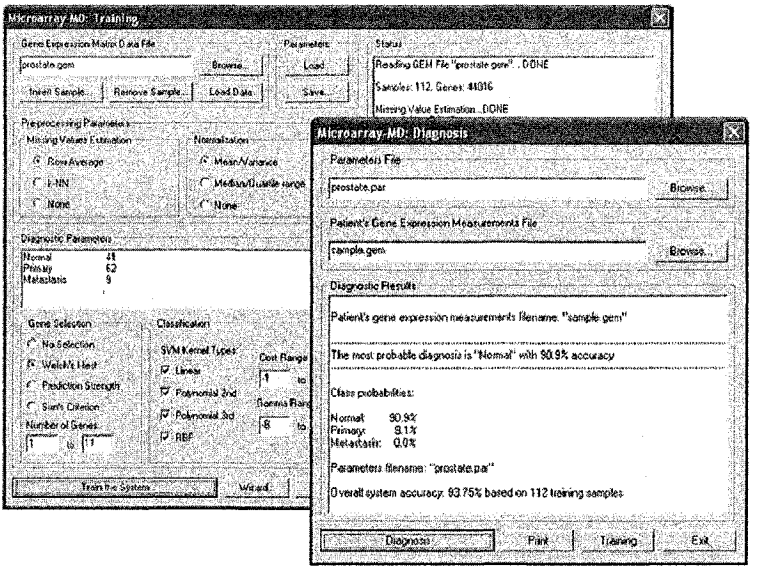

Fig. 1. The system's graphical user interface.

2.1 Pre-processing Unit

The Pre-processing Unit handles the management of missing values as well as the normalization of the gene expression levels. Poor quality in the preparation of the cDNA targets contributes to low quality gene expression measurements, as it affects the mean values and the standard deviation of the spots' intensities, their size and their contrast to the local background areas [14]. Such low quality measurements are usually discarded and missing values appear.

In the Pre-processing Unit of the proposed system we have included a) the row-average method, as it is simple and effective [14] and b) the *k*-nearest neighbours method (*k*-NN) which is more robust than the row-average method but requires more computations [15].

2.2 Diagnostic Unit

The Diagnostic Unit handles medical diagnosis as a multi-class classification problem. It is capable of classifying the input gene expression vectors to N classes noted as ω_i, $i = 1, 2... N$. Each class corresponds to samples acquired from healthy patients, from patients suffering from the same disease or from patients suffering from a subtype of a particular disease. It comprises of N-1 cascading blocks B_j, $j = 1, 2, ... N$-1 as illustrated in Fig. 2.

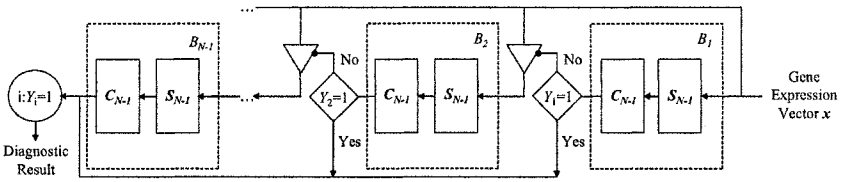

Fig. 2. The Diagnostic Unit.

Each block consists of a Gene Selection Module S_j and a Classification Module C_j. Module S_j uses the output of the Pre-processing Unit as input. Module C_j is autonomously trained with a subset X_j of the available training samples X, where X_j is defined as

$$X_j = \{x \in (\omega_j \cup \omega_h)\},\ \omega_h = \bigcup_{p=j+1}^{N} \omega_p \tag{1}$$

Module S_j selects a subset of τ_j gene expression measurements which best discriminates class ω_j from class ω_h and maximizes the classification performance of the module C_j. Given a test vector x, the module C_j is fed with τ_j gene expression measurements and outputs 1, if $x \in \omega_j$, or -1, if $x \notin \omega_j$. If $x \notin \omega_j$, the next block B_{j+1} will be activated to classify the test vector using the corresponding τ_{j+1} gene expression measurements. Otherwise the classification task terminates and x is assigned to class ω_j. The last block B_{N-1} decides whether $x \in \omega_{N-1}$ or $x \in \omega_N$.

The gene selection modules of the Diagnostic Unit integrate three ranking criteria for the selection of differentially expressed genes (Eqs. 2-4) [16-18] have shown that these criteria can be efficiently used for the identification of differentially expressed genes. These criteria suggest that the genes are ranked in descending order based on the absolute value of the $Z(g)$ statistic for each gene g.

$$Z(g) = \frac{m_g^j - m_g^h}{\sigma_g^j + \sigma_g^h} \tag{2}$$

$$Z(g) = \frac{m_g^j - m_g^h}{\sqrt{\frac{(\sigma_g^j)^2}{n_j} + \frac{(\sigma_g^h)^2}{n_h}}} \tag{3}$$

$$Z(g) = \frac{n_j(m_g^j - m_g)^2 + n_h(m_g^h - m_g)^2}{\sum_{i\in\omega_j}(x_{gi} - m_g^j)^2 + \sum_{i\in\omega_h}(x_{gi} - m_g^h)^2} \quad (4)$$

The (m^j_g, σ^j_g) and (m^h_g, σ^h_g) correspond to the mean and standard deviation of the expression levels of the gene g for the training samples that belong to ω_j and ω_h classes respectively and m_g is the mean expression level of gene g for the entire training set. The x_{gi} is the (g, i) element of the gene expression matrix that corresponds to the expression level of gene g for the sample i. The number of samples belonging to each of the above classes is denoted by n_j and n_h. The τ_j top-ranked genes are selected as they lead to a large between-class distance and a small within-class variance.

The classification module of each block of the Diagnostic Unit implements a binary SVM classifier. SVM training involves a quadratic programming optimization procedure which aims to the identification of a subset of vectors from the available set of training vectors x_i, i=1,2,...n called *support vectors*. These vectors are utilized in the derivation of a separating hypersurface that separates the two classes $y_i \in \{-1, 1\}$, according to the following equation

$$\sum_{\forall i:0<\lambda_i\le c} \lambda_i y_i K(x_i, x) + w_0 = 0 \quad (5)$$

where $0 < \lambda_i < c$ are Lagrange multipliers that correspond to the support vector solutions, c is a positive cost parameter, and $K(x_i, x_j)$ is a kernel function that maps the input space into a high dimensional Hilbert space [19].

In the diagnostic mode of operation, given a test vector x_i, the trained SVM outputs a label Y in accordance with the following formula

$$Y = sign\left(\sum_{\forall i:0<\lambda_i\le c} \lambda_i y_i K(x_i, x) + w_0\right) \quad (6)$$

which designates the class in which an unknown vector x_i belongs to. This information is subsequently used for the derivation of the final diagnostic result.

3 Results

Experiments were conducted on publicly available datasets to evaluate the performance of the proposed system for the diagnosis of diseases. We summarize the results of the application of the proposed system for prostate cancer diagnosis. The prostate cancer dataset used was first studied by Lapointe et al. [20] and it is available from the Stanford Microarray Database [21]. It consists of 112 samples with 44,016 gene expressions spanning three classes, namely 62 primary prostate tumors, 41 normal prostate samples and 9 pelvic lymph node metastases.

The gene expression matrix data file of the prostate cancer dataset was loaded to the system and the structure of the diagnostic unit was determined to two blocks. The first block was assigned to the discrimination of the normal from the joint primary and metastatic samples, while the second block was assigned to the discrimination of primary from metastatic samples.

Comparative classification results were obtained by running the experiments also using the well established one-vs-one SVM combination scheme [22]. During the training mode of the SVM-based classification schemes (Cascading and one-vs-one) the kernel functions tested were the linear, the 2nd and 3rd order polynomial and the RBF [19]. The ranges of the training parameters considered were 2^{-5} to 2^{15} for the cost parameter c and 2^{-15} to 2^{3} for the γ parameter. The best parameters that maximize the performance for the cascading and the one-vs-one schemes were determined automatically using grid search. The order of the blocks in the cascading model was determined based on the histopathological sub-classification of carcinomas [20]. A range of one to 11 genes was considered in the gene selection process.

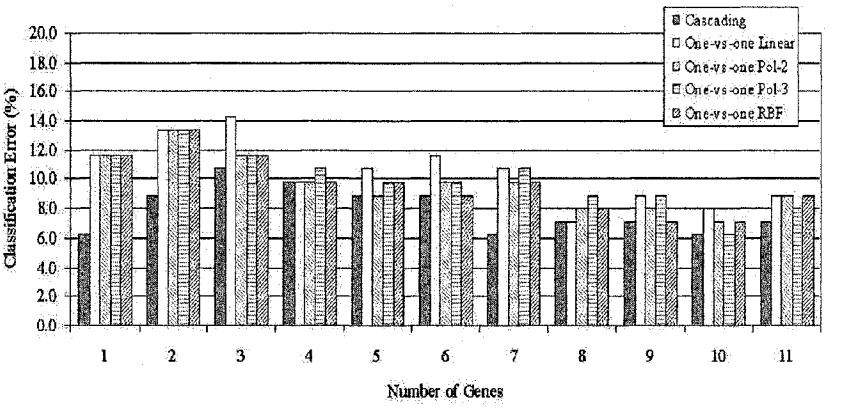

Fig. 3. Diagnostic accuracy of the proposed system (cascading SVM architecture) compared with the standard one-vs-one SVM combination scheme, for various numbers of input genes.

In all cases, the proposed architecture results in lower or comparable classification error rates with the one-vs-one scheme. The minimum classification error obtained for the prostate cancer dataset reached 6.3% in three cases using one, seven or ten genes. The same classification error rate was obtained by one-vs-one scheme using 3rd order polynomial kernel and ten genes. So, the proposed architecture has an advantage over these schemes as it is capable of providing a better or comparable performance using fewer genes. The upper classification error bound of the cascading SVMs architecture is 10.7%, whereas this bound increases to 14.3% for the one-vs-one scheme.

The results are presented in Fig. 3. The diagram shows that the proposed architecture leads to lower classification error rates compared to the one-vs-one classifier using one to five genes. The one-vs-one SVM combination scheme with linear kernel resulted in the lowest classification error using six input genes.

The classification errors achieved by the proposed system on other publicly available datasets were also low. These include colon cancer (9.7%), and lung cancer (1.5%) datasets [23][24].

4 Conclusions

We presented a biomedical software system capable of supporting medical diagnosis using gene expression data produced by microarray experiments. The major contribution of the proposed system in the process of medical diagnosis is that it offers to the physicians substantial molecular-level information by exploiting gene expressions. The gene expression measurements are pre-processed and subsequently used for the classification of the corresponding samples in two or more categories depending on their pathology.

The proposed system has been tested on various publicly available microarray datasets, including those provided by Stanford Microarray Database [21]. In most cases the overall diagnostic accuracy it provides, exceeds 90%. Its high accuracy was avouched in this paper by demonstrating its application for prostate cancer diagnosis. It can be concluded that the proposed system can be used as a valuable diagnostic aid by physicians and with the decreasing cost of microarrays it could potentially be used in everyday medical practice.

Moreover the cascading SVM combination scheme provides low classification error rates which are comparable and in most cases lower than the rates obtained by the one-vs-one SVM combination scheme especially when a small number of genes is involved. The proposed architecture utilizes N-1 classifiers whereas the one-vs-one SVM combination scheme utilizes $N(N\text{-}1)/2$ classifiers and the one-vs-one SVM combination scheme utilizes N classifiers.

Currently, the approach followed for the system to learn from new training data involves discarding the existing classifier, combining the old and the new data and training a new classifier from scratch using the aggregate data. Within our prospects is the enhancement of the proposed system by incorporating an incremental approach to SVM learning that will allow efficient on-line training without losing prior knowledge from additional datasets that will later become available.

Acknowledgment

This research was funded by the Operational Program for Education and Vocational Training (EPEAEK II) under the framework of the project "Pythagoras - Support of University Re-search Groups" co-funded by 75% from the European Social Fund and by 25% from national funds. We would also like to thank the anonymous reviewers for their helpful comments and suggestions.

References

1. Do, K.-A., Nikolova, R., Roebuck, P., Broom, B.: GeneClust, http://odin.mdacc.tmc.edu /~kim/geneclust/, accessed Nov. 2004
2. Hastie, T., Tibshirani, R., Eisen, M. B, Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein D., and Brown, P.: 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Gen. Biol. 1 (2000) 0003.1-0003.21
3. Li, C., Wong, W. H.: Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. PNAS 98 (2001) 31-36
4. Peterson, L.E.: CLUSFAVOR 5.0: hierarchical cluster and principal-component analysis of microarray-based transcriptional profiles. Gen. Biol. 3 (2002) 0002.1-0002.8

5. Sturn, J. Quackenbush, Z. Trajanoski: Genesis: cluster analysis of microarray data. Bioinformatics 18 (2002) 207-208
6. Colantuoni, C., Henry, G., Zeger, S., Pevsner, J.: SNOMAD (Standardization and NOrmalization of MicroArray Data): web-accessible gene expression data analysis, Bioinformatics 18 (2002) 1540-1541
7. Saal, L. H., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, Å., Peterson, C.: BioArray Software Environment: A Platform for Comprehensive Management and Analysis of MicroarrayData. Gen. Biol. 3 (2002) 0003.1-0003.6
8. Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V., Quackenbush, J.: TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34 (2003) 374-378
9. Gentleman, R., Rossini, R., Dudoit S., Hornik K.: The Bioconductor FAQ, (2003) official URL. http://www.bioconductor.org/
10. Yang, S., Murali, T. M., Pavlovic, V., Schaffer, M., Kasif, S.: RankGene: identification of diagnostic genes based on expression data. Bioinformatics. 19 (2003) 1578-1579
11. Xu, D., Olman, V., Wang, L., Xu, Y.: EXCAVATOR: a computer program for efficiently mining gene expression data. Nucleic Acids Research 31 (2003) 5582-5589
12. Toyoda T., Konagaya, A.: KnowledgeEditor: a new tool for interactive modeling and analyzing biological pathways based on microarray data. Bioinformatics. 19 (2003) 433-434
13. Pieler, R., Sanchez-Cabo, F., Hackl, H., Thallinger G.G., Trajanoski, Z.: ArrayNorm: comprehensive normalization and analysis of microarray data. Bioinformatics. 20 (2004) 1971-1973
14. Zhang, W., Shmulevich, I., (ed.), Computation and Statistical Approaches to Genomics, Kluwer Academic Publishers, Boston, (2002)
15. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshiran, R., Botstein D., Altman, R.B., Missing value estimation methods for DNA microarrays. Bioinformatics 17 (2001) 520-525
16. Pan, W., A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics. 18 (2002) 546-554
17. Golub, T.R. et al.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science. 286 (1999), 531-537
18. Sun, M., Xiong, M.: A mathematical programming approach for gene selection and tissue classification. Bioinformatics 19 (2003) 1243-1251
19. Vapnik, V.: Statistical Learning Theory, John Will and Sons, New York, (1998)
20. Lapointe, J., Li, C., Higgins, J.P., Van de Rijn, M., Bair, E., Montgomery, K. et al. : Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc. Nat. Acad. Sci. 101 (2004) 811-816
21. Stanford Microarray Database, http://genome-www5.stanford.edu., accessed Nov. 2004.
22. Hsu C.W., Lin, C.J., A comparison of Methods for Multiclass Support Vector Machines, IEEE Trans. Neural Networks, 13 (2002), 415–425
23. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A. J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Nat. Acad. Sci. 96 (1999) 6745–6750.
24. Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa P., et al.:, Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Nat. Acad. Sci. 98 (24) (2001) 13790-13795