2011 International Conference on Environment and BioScience
IPCBEE vol.21 (2011) © (2011) IACSIT Press, Singapore

DMCA: A Combined Data Mining Technique for Improving the Microarray Data Classification Accuracy

Dina A. Salem[1+], Rania Ahmed A. A. Abul Seoud[2] and Hesham A. Ali[3]

[1] Dept. of Electronics Engineering-Faculty of Engineering-Misr University for Science and Technology, Egypt

[2] Dept. of Electrical Eng.- Comm. and Electronics Section-Faculty of Eng - El Fayoum University- Fayoum, Egypt

[3] Computer Engineering System Dept-Faculty of Engineering-Mansoura University, Egypt

Abstract. Data mining techniques are important to sift through the huge amount of gene expression values in microarrays resulting in valuable biological knowledge. An important example is classifying cancer samples, which is crucial to biologists for cancer diagnosis and treatment. In this paper we propose the DMCA technique in which the main objective is reducing the number of genes needed for accurate classification. The proposed technique is a combination of two feature selection techniques, f-score and entropy-based, and a powerful classifier, Support Vector Machines. DMCA achieved promising results and is characterized by being flexible in all of its stages. When applied to two public microarray datasets, DMCA succeeded in reducing the number of gene expression values needed to classify a sample by 71.29% and guaranteed reliable classification accuracy.

Keywords: data mining, bioinformatics, microarrays, classification, gene selection, f-score technique

1. Introduction

Data is tremendously growing in all life aspects resulting in mountains of data. Mining these mountains using powerful data analysis tools is important to obtain the contained valuable information needed to present decision making solutions. Data Mining is the automated process of analyzing data from different perspectives to extract previously unknown, comprehensible, and actionable information hidden in large data repositories and using it to make crucial decisions [1, 2]. One of the important data mining methods is classification. Classification is a supervised machine learning technique works on classifying a new data item into a predefined class [1], and so it will be the main concern of our work. Data mining is efficiently applied to almost all computerized fields of life resulting in powerful and reliable solutions for thousands of problems [3]. One of the well known and most important data mining applications is bioinformatics. Bioinformatics can be defined as the application of computer technology to the management of biological information. In bioinformatics, data mining has a primary goal in increasing the understanding of biological processes. Some of the grand areas of research in bioinformatics are analysis of gene expressions and mutations in cancer [4]. Cancer databases and gene expression values are the data used in this paper and is extracted from the emerging microarray technology. High-throughput microarray technology is a hybridization procedure that enables the simultaneous measurement of the abundance of tens of thousands of gene-expression levels from many different samples on a small chip [5].

The large amount of data in microarrays makes it in a deep need for data mining. Microarray data is mainly used in cancer diagnosis and prognosis where it's well known that the early diagnosis of cancer and determining its type is very helpful in its treatment. Data mining classification techniques are very suitable to

[+] Corresponding author. Tel.: + 2020101662398.
E-mail address: dena.salem@mail.com.

address this issue where new samples can be classified into two or more predefined cancer classes using the gene expression values. Microarray data is characterized by its high dimensionality which means that the number of samples (always less than 100) is not proportional to the number of genes (always thousands). This explains the need for a feature selection technique before entering the data into the classifier. The feature in the microarray data is the gene and the feature selection is renamed to be a gene selection [6]. Then, the process of classifying microarray data merges two main steps; implementing an effective gene selection technique and choosing a powerful classifier. These two steps will form the workflow of this paper trying to go through each step details and validating its outputs.

The remainder of this paper is organized as follows; Section 2 reviews briefly some of the recent work published in the area of classification of cancer using microarray gene expression values. Section 3 introduces and describes the general scheme of our proposed combined data mining technique. Results of the proposed technique are presented in section 4. Section 5 analyzes these results. Finally, section 6 concludes the paper.

2. Related Work

A lot of research has addressed the topic of the classification of the microarray data by using different gene selection methods with different classifiers. A generic approach to classifying two types of acute leukemias was introduced in Golub et al. [7]. Two other systems used for classifying the same microarray dataset was by blending of Support Vector Machine as a classifier, once with Locality Preserving Projection technique (LPP) and the other with F-score ranking feature selection technique [8, 9]. SVM is used again by Moler *et al.* but this time combined with a naive Bayesian model for classifying the colon adenocarcinoma tissue specimens labeled as tumor or nontumor dataset for the first time [10]. The two previous datasets were used by P. Yang and Z. Zhang to validate their two proposed systems (GADT, GANN) which combined the genetic algorithm (GA) for gene selection with two classifiers; Decision Tree and Neural Network [11]. J. Zhang and H. Deng chose their reduced set of genes by first carrying a gene preselection using a univariate criterion function and then estimating the upperbound of the Bayes error to filter out redundant genes from remaining genes derived from gene preselection step. To validate their system they used two classifiers; k-nearest neighbor (KNN) and SVM on five datasets [12].

2.1. Problem Formulation

Most of the previous work uses only one gene selection technique to reduce the original number of genes. This means that they take a single gene criterion into their consideration according to the chosen technique. Sometimes this is not sufficient to obtain the highly informative genes. Thus, it's a challenge to design a classification system which is capable of classifying new samples using a smaller highly informative gene subset of the original set of genes and, at the same time without resulting in any misclassifications. For this purpose, we proposed DMCA technique which combines two gene selection techniques for reducing the number of involved genes. Each one of them selects the informative genes according to a different criterion. DMCA is trained by a set of samples with known predefined classes. DMCA technique can identify and record the classes of a set of unclassified samples using the reduced gene subset. The efficiency of the DMCA technique can be evaluated by measuring the classification accuracy.

3. Methodology

The DMCA technique receives preprocessed high dimensionality microarray dataset as its input. The technique first step is reducing the total number of genes in the input dataset to a smaller subset using F-score and entropy ranking techniques as a combined gene selection technique. Then the reduced data trains the chosen Support Vector Machine classifier until it's completely learned by the help of the cross-validation. At this point we can measure the DMCA train classification accuracy (train CA = number of correctly classified samples divided by total number of train samples). Then comes the last step where the system is ready to receive new samples (which are not used in the training of the SVM classifier) and classify them using only the reduced subset of the genes and once again measure the test classification accuracy (test CA = number of correct classified samples divided by the total number of test samples). The workflow of the DMCA technique is shown in Fig.1.

3.1. Microarray Gene Expression Datasets

When working with the DMCA technique, any used dataset must be split into two sub-datasets; a training dataset which the classifier uses to learn and form its learned structure and, a test dataset to see the effectiveness of the proposed system. DMCA technique works on two public datasets and can be extended to classify other datasets. One dataset is the leukemia dataset which was first classified by Golub et al.in 1999 into two classes; Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML) [7]. The other dataset is the lymphoma dataset which was classified by Shipp et al. in 2001 into two classes; Diffuse Large B-Cell Lymphoma (DLBCL) and Follicular Lymphoma (FL) [13]. Each sample in both datasets has expression patterns of 7129 genes measured by the Affymetrix oligonucleotide microarray. The two datasets are available and downloaded from Broad Institute of MIT and Harvard website (www.broadinstitute.org). Table 1 contains the details of the two datasets.

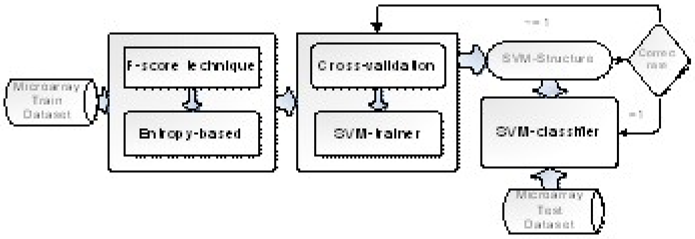

Fig. 1: DMCA technique workflow.

Table 1: Used datasets details.

Dataset	*Classes*	*Genes*	*Total samples*	*Train samples*	*Test samples*
Leukemia	AML,ALL(2)	7129	72	38	34
Lymphoma	DLBCL,FL(2)	7129	77	40	37

3.2. Gene Selection

Cancer microarray data usually consists of a few hundred samples with thousands of genes as features. Classification of data in such a high dimensional space is impossible as this may lead to over fitting, in addition to the ultimate increase in the processing power and time [14]. This gives rise to the need of the gene selection techniques which aim to find a subset of highly informative and relevant genes by searching through the space of features. These techniques fall into three categories; marginal filters, wrappers and embedded methods. Marginal filter approaches are individual feature ranking methods. In a wrapper method, usually a classifier is built and employed as the evaluation criterion. If the criterion is derived from the intrinsic properties of a classifier, the corresponding feature selection method will be categorized as an embedded approach [15].

Filter methods are characterized over the two other types by being powerful, easy to implement and is a stand-alone technique which can be further applied to any classifier. They work on giving each gene a score according to a specific criterion and choosing a subset of genes above or below a specified threshold. Thus, they remove the irrelevant genes according to general characteristics of the data [16]. In our technique, instead of using only one filter technique, we use a combination of two efficient techniques; the F-score and the entropy-based. The F-score ranks the genes twice; one time according to the two classes mean difference for each gene and choosing the top 250 genes then, according to the Signal-to-Noise ratio (SNR) criterion and choosing the highest 200 genes. So it can identify the genes whose expression shows great change in both the classes [9]. The entropy-based technique ranks the subset of genes resulting from the F-score technique according to their entropy value and chooses the first 100 genes. The combined technique is implemented in MATLAB 7.10.0 (R2010a).

3.3. Cross-validation

Cross-Validation (CV) is very helpful in evaluating and comparing learning algorithms. It is a statistical technique used during the training process of the classifier where its task is to divide the train dataset into two segments; one is used for training and the other is used for validation. The training and validation sets must cross-over in successive rounds such that each sample has a chance of being validated against. CV is carried out in different forms where the most general form is the k-fold cross-validation. One special form of the k-fold CV is the leave-one-out-cross-validation (LOOCV) where it uses one sample for test and all other samples for training [17]. No agreement exists on a common value of k for CV in microarray data classification. As the number of samples used in the training of the classifier usually doesn't exceed several tens, we consider four values of k in our study (1, 2, 5, 10). Cross-validation is available in the bioinformatics toolbox in MATLAB 7.10.0 (R2010a).

3.4. Support Vector Machine

Support Vector Machines (SVMs) have been widely used in the recent years in the field of computational biology due to their high accuracy and their flexibility in modeling diverse sources of data. They are mainly used in binary classification and regression. They are very suitable for classifying microarray gene expression data [18]. The SVM is partitioned into two parts; the SVM-trainer and the SVM-classifier.

- SVM-trainer: In this section SVM uses the train dataset to construct a hyperplane to separate two sets of data points (samples). It solves an optimization problem to reach the maximum margin. The maximum margin is the largest distance from the hyperplane to the nearest data points. Data points fall on this margin are called support vectors. This can be easily achieved for linear separable data points. Otherwise, SVM uses the kernel functions to map the non-linear separable samples into the feature space. Different kernel functions include; Gaussian, polynomial, and RBF. The output of this stage is the SVM-structure.
- SVM-classifier: In this section SVM uses the SVM-structure to classify the test data into the predefined classes. As the CV and SVM parameters are accurately chosen, as the classification accuracy of the test samples increases. The SVM-trainer and SVM-classifier are available in the bioinformatics toolbox in MATLAB 7.10.0 (R2010a).

The cross-validation coupled with the SVM-trainer runs several times in a continuous loop until reaching maximum train classification accuracy (correct rate=1). If the correct rate does not reach the value 1, the loop is manually stopped recording the highest value of the correct rate. Achieving highest train classification accuracy forms the optimum SVM structure which in turn leads to minimum misclassifications for the test samples. The SVM structure is the output of the SVM trainer. It contains all the information needed for the SVM classifier to classify the new test samples.

4. Results

The large number of attributes involved in each of the DMCA technique stages and the variety of values each attribute can take makes the evaluation process exhaustive. But to fully evaluate our system in a comprehensive way, we take into our consideration the effect of every change and record the classification accuracy (CA) twice every time. Once for the training stage, giving it the name train classification accuracy (train CA) and again for the test stage with the name test classification accuracy (test CA). The CA is not the only important issue to evaluate the technique, the number of genes used for training and test has the same importance. Thus, we recorded the train CA and test CA for a subset of 200 genes (table 2, 3) and again for a subset of 100 genes (table 4, 5). The 200 genes subset is the result of the f-score gene selection technique only without combining the entropy- based technique. While, the 100 genes subset is the result of our combined gene selection technique. Tables 2 and 3 emphasize the effect of two cross-validation methods, the LOOCV and the K-fold with three k values (2,5,10), on the train CA and test CA for both datasets used. SVM is applied with linear kernel function. Applying polynomial and Gaussian kernel functions decrease the CA. So, we ignore recording the results when using kernel functions for SVM.

Table 2: Train/Test CA on 200 genes subset.

dataset	LOOCV	k-fold CV		
		K=2	K=5	K=10
Leukemia	1/0.97	1/1	1/0.94	0.9737/0.94
lymphoma	1/0.9459	0.975/0.9459	0.975/0.9459	0.925/0.9459

Table 3: Train/Test CA on 100 genes subset.

dataset	LOOCV	k-fold CV		
		K=2	K=5	K=10
Leukemia	1/0.9706	1/1	1/1	1/0.9706
lymphoma	1/0.9459	0.95/0.9459	0.975/0.9189	0.95/0.973

For further evaluation of the DMCA technique, we compared its results with some published papers which are previously mentioned in the related work section. In this comparison we take two attributes into our consideration; the test CA and the reduced number of genes used in the classification process. As DMCA was applied to two different datasets, we compared the results for each dataset with the results from published papers dealing with the same dataset, separately from the other. For the leukemia dataset, DMCA results are compared with [8, 9, 11]. For the lymphoma datasets, the comparison was with [12, 16]. Our technique shows the highest test CA for the two datasets over the other techniques with a considerable number of genes. Figure 2, 3 show the compared results where name of the technique used and number of involved genes (between parentheses) are written on the horizontal axis, and the test CA on the vertical axis.

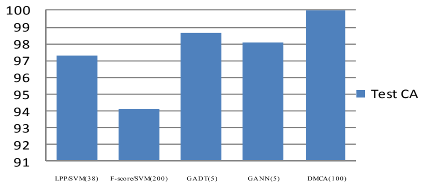

Fig. 2: Compared results of leukemia dataset.

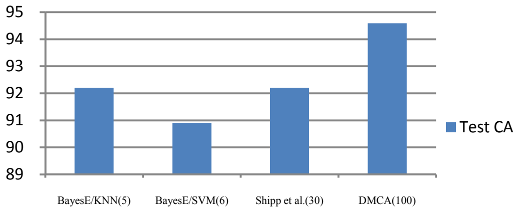

Fig. 3: Compared results of lymphoma dataset.

5. Analysis

As seen from the previous DMCA results, the highest train and test classification accuracies for leukemia dataset is 1(means all train and test samples are classified correctly) when using k-fold Cross Validation with k=2. Also the highest train and test classification accuracies of lymphoma dataset are 1 and 0.9459 which means all the training samples are classified correctly and only two test samples are misclassified. This is achieved when applying leave-one-out-cross-validation. These great results occur in both reduced datasets. This means that the chosen entropy-based gene selection technique is of a great value as it reduced the number of genes needed to classify a microarray sample to 50% of the number results from applying f-score technique only. Also we noticed that applying different kernel functions didn’t enhance the classification accuracy but sometimes reduced it. However, when applying DMCA on other microarray gene expression datasets, a need to kernel functions may arise to increase the CA.

6. Conclusion

In this paper, we proposed the DMCA technique, a technique for classifying cancer samples using microarray gene expression datasets. The main target of the technique is to get the highest accuracy when classifying the samples using a small subset of informative genes. A combination of two gene selection techniques is introduced to solve the problem of the microarray high dimensionality. This combined technique presents high performance as it reduces the number of genes by 71.29%. SVM was chosen for classification as it is a very efficient binary classification technique and always gives good results by attenuating its variety of attributes. DMCA technique was applied to two public microarray datasets, leukemia dataset and lymphoma dataset. It shows a maximum accuracy on leukemia dataset without any misclassifications and a very small error rate on lymphoma dataset equals to 0.0541. Thus, DMCA technique reaches its main objective as it classified the test cancer samples with high classification accuracy (100% on leukemia dataset and 94.59% on lymphoma dataset) using only a set of 100 genes. Comparison with others verifies the efficiency of the proposed technique. The DMCA is a powerful flexible technique which can be adapted to any microarray gene expression dataset. This technique can easily be extended to integrate any other gene selection technique or any other classifier.

7. References

[1] D. T.Larose. Discovering knowledge in Data: An Introduction to Data Mining: *John Wiley & Sons, Inc*. 2005.

[2] E. Simoudis. Reality check for data mining. *IEEE Expert*. 1996, 26-33.

[3] J. Han, M. Kamber. Data Mining:Concepts and Techniques, 2nd edn: *Morgan Kaufmann Publishers*. 2006.

[4] K. Raza. Application Of Data Mining In Bioinformatics. *Indian Journal of Computer Science and Engineering.* 2010, 1(2):114-118.

[5] C. Kong, J. Yu, F. Minion,K. Rajan. Identification of Biologically Significant Genes from Combinatorial Microarray Data. *ACS Combinatorial Science.* 2011.

[6] H. Ong,N. Mustapha,M. Sulaiman. Integrative Gene Selection for Classification of Microarray Data. *Computer and Information Science (CCSE).* 2011, 4(2):55-63.

[7] T. Golub et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.* 1999,286:531–537. doi: 10.1126/science.286.5439.531.

[8] J. Salome, R. Suresh. An Effective Classification Technique for Microarray Gene Expression by Blending of LPP and SVM. *Medwell Journals : Asian Journal of Information Technology.* 2011, 10(4):142-148.

[9] K. Seeja, Shweta. Microarray Data Classification Using Support Vector Machine. *International Journal of Biometrics and Bioinformatics (IJBB).* 2011, 5(1):10-15.

[10] E. Moler, M. Chow, I. Mian. Analysis of molecular profile data using generative and discriminative methods. *Physiological Genomics.*2000, 4(2):109-126.

[11] P. Yang, Z. Zhang. Hybrid Methods to Select Informative Gene Sets in Microarray Data Classification. In: M. Orgun, J. Thornton (eds.). *Proc of Australian Conference on Artificial Intelligence.* Verlag Berlin Heidelberg: 2007,810-814.

[12] J. Zhang, H. Deng. Gene selection for classification of microarray data based on the Bayes error. *BMC Bioinformatics.* 2007, 8(1):370.

[13] M. A. Shipp et al.. Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene Expresssion Profiling And Supervised Machine Learning. *Nature Medicine* 2001, 8(1):68-74.

[14] E. B. Huerta, B. Duval, J.-k. Hao. A hybrid ga/svm approach for gene selection and classification of microarray data. In: *EvoWorkshops, LNCS 3907. 2006*,pp. 34-44.

[15] Guyon, A. e. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research.* 2003, 3:1157-1182.

[16] Y. Wanga, I. V. Tetkoa, M. A. Hallb, E. Frankb, A. Faciusa, K. F. X. Mayera, H. W. Mewesa. Gene selection from microarray data for cancer classification. *Computational Biology and Chemistry.* 2005, 29(1):37-46.

[17] M. Saeedmanesh, T. Izadi, E. Ahvar. HDM: A Hybrid Data Mining Technique for Stock Exchange Prediction. In: *International MultiConference of Engineers and Computr Scientists (IMECS).* Hong Kong: 2010.

[18] Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf, G. Rätsch. Support Vector Machines and Kernels for Computational Biology. *PLoS Comput Biol* 2008,4(10).e1000173. doi:10.1371/journal.pcbi.1000173