High Dimensional Microarray Data Classification Using Correlation Based Feature Selection

Abid Hasan
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Gazipur, Bangladesh
aabid@iut-dhaka.edu

Md. Akhtaruzzaman Adnan
Department of Computer System and Communication
University of Technology Malaysia (UTM)
Skudai, Johor Bahru, Malaysia
adnanradowan@yahoo.com

Abstract— **Analyzing DNA microarray data pose a serious challenge because of their large number of features (genes) and relatively small number of samples. Extracting features, those have predictive capability for classifying these huge datasets demands appropriate approaches like feature reduction and identifying optimal set of genes. In this paper along with conventional statistical methods like filtering the dataset to reduce the number of features, one additional approach of evaluating correlation between the classes for each feature is performed. Proposed approach yields higher classification accuracy for both Acute Lymphoblastic (ALL) and High Grade Glioma cancer dataset than using only traditional statistical filtering methods.**

Keywords- DNA microarray data, correlation, feature selection, classification

I. INTRODUCTION

Microarray shows enormous potential in medical science in recent years as large scale measurement of activities of genes by measuring expression level of thousands of genes simultaneously [1, 2, 3, and 4]. Computational complexity makes these data difficult to represent in a meaningful way for the researchers. In addition it also faces the challenges of feature selection, noises, missing values, high dimensionality background and special effects [5]. Among these enormous numbers of genes only few of them show significant discriminatory ability as features for high dimensional microarray dataset classification. Especially the curse of dimensionality can be addressed by applying feature selection and dimensionality reduction methods.

Feature selection is performed to discard uncorrelated and redundant features (genes) in order to get the most discriminative genes. Several researches use dimensionality reduction methods assuming that many genes are unnecessary and irrelevant to constructing the classification model [6, 7, 8 and 9]. Redundancy of genes in these dataset also doesn't play any significant role for model building. [1]. Moreover *trustworthiness* and *Minimum redundancy and maximum relevance* (MRMR) is also used for filtering data for feature reduction [5, 7].

In this paper the idea of feature selection using traditional filtering methods has been extended by using correlation based ranking of the genes where the correlation is computed between the two subtypes of genes that can cause cancer. Here the proposed method is applied on Acute Lymphoblastic Leukemia (ALL) subtypes: B-cell ALL, T-cell ALL and High Grade Glioma subtypes: classic and non-classic glioblastoma and classic and non-classic oligodendroglioma cancer data. Results of classifying these cancer data on two different settings of the proposed method are compared for experimental evaluation where one approach is only using the traditional selected filtering methods and the proposed method of feature selection.

II. FEATURE SELCTION

In the process of computing correlation between the two subtypes of datasets, two vector x and y is taken where these vectors contain the sample mean for each gene.

$$x = [S_{x1}, S_{x2}, ..., S_{xn}]\,;\; y = [S_{y1}, S_{y2}, ..., S_{yn}]$$

where $S_{xi} = \frac{1}{m}\sum_{j=1}^{m} s_{xj}$ and $S_{yi} = \frac{1}{m}\sum_{j=1}^{m} s_{yj}$

here n is the number of genes in the dataset and m is the number of samples.

s_{xi} is the expression values of each sample in one of the subtypes of dataset and s_{yi} is the expression values of other subtype. S_{xi} and S_{yi} is the arithmetic mean of expression values of each gene of all the samples in their corresponding subtypes respectively. Vector V contains the variation of each gene by calculating the difference between the each element of x with $\overline{X}$ and y with $\overline{Y}$ where

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Now, $V = \left|x_i - \overline{X}\right| * \left|y_i - \overline{Y}\right|$ where $i = 1...n$

The correlation between the two subtypes of data can be obtained by

$$C = V_i / (\sigma_x * \sigma_y) \quad (1)$$

where σ_x and σ_y are the standard deviation of two subtypes of the datasets respectively.

Each gene is associated with their correlation value *C* and the dataset is ordered based on their correlation value. In the next step those genes are selected who has correlation values higher than the certain threshold. The threshold is defined by a value that indicates minimum correlation for the features. Here the minimum value is set to 1. Next this list of genes is considered to be the input to a conventional feature selection algorithm. A number of conventional feature selection algorithms are selected for this step.

III. METHODOLOGY

One of the two datasets used for experimentation in this paper is ALL dataset which was used in Sabina et al. [10] containing 128 samples who had Acute Lymphoblastic Leukemia (ALL). Among them 95 of them had B-cell ALL and 33 of them had T-cell ALL. The experiment was performed on a HGU95AV2 Affymetrix single channel microarray chip that contains 12,625 genes.

The second dataset for our experiment is High Grade Glioma dataset that was used in Nutt et al. [11] to determine the expression of approximately 12,000 genes using microarray analysis over 50 gliomas. Among them 26 samples are of classic and non-classic glioblastomas and 24 of them are classic and non-classic oligodendrogliomas. For this experiment we used Affymetrix U95AV2 GeneChips.

Data was preprocessed and normalized using robust multi-array average (RMA) expression measure which is and expression measure obtained by three steps: convolution background correction, quantile normalization and a summarization based on a multi-array model fit robustly using median polish algorithm [12, 13]. Correlation measurement and other filtering were implemented in R language [14]. Attribute selection method used here was implemented using Weka workbench [15]. After we obtain the potential features (genes) for classifying the dataset, two of the popular decision tree classifiers C4.5 and CART are used for classification in Weka [16].

In the first step, correlation is calculated between the two subtypes of datasets using (1). Through repeated experiments, the threshold for each of the datasets was determined. In the next step, non-specific filtering was performed. In each dataset the gene must have an expression level that is greater than log(200) in at least 25% of the samples. It must have a median expression level that is greater than log(300) and it must have an IQR (Inter-quartile Range) that is larger than 0.5. The filtering method used here is *ttest*, used in Yu et al. [17]. The genes those who are not differentially expressed among the two subtypes of dataset are removed by this filtering method. *ttest* returns a function of one argument with bindings for *cov* and *p* (i.e. the false discovery rate). When evaluated, this function performs a *t-test* using the *cov* as the covariant. It returns TRUE if the *p* value for the difference in their means is less than *p*. Here, $p = 0.01$ is used.

After filtering, the reduced set of genes for both ALL dataset and High Grade Glioma dataset was obtained. Considerable reduction in the number of genes was observed for both approaches where the genes were ranked based on their correlation value, and then applies filtering and other approach where only filtering is used. Next, a conventional attribute selection method was applied to both gene lists resulted from previous step. Supervised attribute filter, *cfsSubsetEval*, in Weka to evaluate the discriminative quality of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. From this set of genes, the list of genes having high correlation with the class and low inter-correlation were selected as features. After performing this step the total number of genes as features is reduced to 27 for the list of genes which was ranked based on their correlation value and 36 where this method was not used for ALL dataset and for High Grade Glioma dataset the number of features is reduced to 11 and 19 for two different approaches mentioned.

IV. RESULTS AND DISCUSSION

The result of the experiment on the two datasets confirms that the microarray dataset classification accuracy is higher in the case where our proposed method is used along with the filtering rather than the other approach where only filtering is applied (see Table I).

TABLE I. CLASSIFICATION ACCURACY USING TWO DIFFERENT APPROACHES ON TWO DATASETS

Dataset	Classification Algorithm	Accuracy (%)	Feature selection methodology
ALL Dataset	C4.5	98.4375	CBRG[a]
		96.0938	Without CBRG
	CART	96.875	CBRG
		96.875	Without CBRG
High Grade Glioma Dataset	C4.5	68.00	CBRG
		64.00	Without CBRG
	CART	80.00	CBRG
		70.00	Without CBRG

[a] Correlation Based Ranked Gene

In both the datasets the accuracy of classifying the dataset is higher with the features (genes) those are ranked based on their correlation value. Although for ALL dataset both the approaches gives the same accuracy rate but the confusion matrix (see Table II) shows the difference between the two approaches of classifying for both the datasets.

It is evident from the above mentioned tables (see Table I and Table II) that the classification accuracy is considerably higher for both the database where the features (genes) are selected using the proposed method.

TABLE II. CONFUSION MATRIX

Dataset	Classifier	Feature selection method	Confusion matrix		Classified As
			x	y	
ALL Dataset	C4.5	CBRG	93	2	x = B
			0	33	y =T
		Without CBRG	94	1	x = B
			4	29	y =T
	CART	CBRG	91	4	x = B
			0	33	y = T
		Without CBRG	94	1	x = B
			3	30	y = T
High Grade Glioma Dataset	C4.5	CBRG	17	11	x = G
			5	17	y = O
		Without CBRG	19	9	x = G
			9	13	y = O
	CART	CBRG	22	6	x = G
			4	18	y = O
		Without CBRG	20	8	x = G
			7	15	y = O

V. CONCLUSION

Reducing the number of features from a large feature space for the purpose of classification is one of the most challenging tasks. Finding the valuable genes as features which has high correlation with the class and low inter-correlation is significant for classifying huge dataset like the two cancer dataset used. It is seen form the experiment that selecting features from a list of genes where the genes are first ranked based on their correlation value gives better feature space rather than not using the approach. Moreover with lesser number of features (genes) requiring for classification with a higher accuracy is another advantage of this proposed method.

REFERENCES

[1] T. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", *Science* 286 (1999)J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] S. Mukherjee, P. Tamayo, D. K. Slonim, A. Verri, T. R. Golub, J. P. Mesirov, T. Poggio, "Support vector machine classification of microarray data", *AI memo 182. CBCL paper 182.Technical report, MIT (2000)* can be retrieved from ftp://publications.ai.mit.edu.K. Elissa, "Title of paper if known," unpublished.

[3] G. Hardiman, "Microarray technologies - an overview". *Pharamacogenomics 3* (2002) 293–297.

[4] M. Schena, "Microarray Biochip Technology", *BioTechniques Press,Westborough*, M.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[5] F. A. Ubaudi, P. J. Kennedy, D. R. Catchpoole, D. Guo, S. J. Simoff, "Microarray Data Mining: Selecting Trustworthy Genes with Gene Feature Ranking" *Data Mining for Business Applications* 2009, II, 159-168, DOI: 10.1007/978-0-387-79420-4_11

[6] A. L. Blum, P. Langley, "Selection of relevant features and examples in machine learning". *Artificial Intelligence* 97 (1997) 245–271.

[7] C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data", *NERSC Division Lawrence Berkeley National Laboratory, University of California, Berkeley, CA, 94720, USA*

[8] B. Efron, R. Tibshirani, V. Goss, G. Chu, "Microarrays and their use in a comparative experiment", *Technical report, Stanford University (2000)*

[9] G. H. John, R. Kohavi, K. Pfleger, "Irrelevant features and the subset selection problem", *Eleventh International Conference (Machine Learning),* Kaufmann Morgan (1994) 121–129.

[10] C. Sabina, X. Li, R. Gentleman, A. Vitale, K. S. Wang, R. Mandelli F Foá, J. Ritz, "Gene Expression Profiles of B-lineage Adult Acute Lymphocytic Leukemia Reveal Genetic Patterns that Identify Lineage Derivation and Distinct Mechanism of Transformation." *Clin Cancer Res*. 2005 Oct 15;11(20):7209-19.

[11] C. L. Nutt, D. R. Mani, Rebecca A. Betensky, P. Tamayo, J. Gregory Cairncross, C. Ladd, U. Pohl, C. Hartmann, M. E. McLaughlin, T. T. Batchelor, P. M. Black, A. von Deimling, S. L. Pomeroy, T. R. Golub and D. N. Louis. "Gene Expression-based classification of malignant gliomas correlated better with survival than hitological classification" *Cancer Res* April 1, 2003 63; 1602

[12] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, T. P. Speed "Exploration Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data". *Biostatistics*. Vol. 4, Number 2: 249-264 (2003)

[13] J. D. Emerson and D. C. Hoagli, "Analysis of two-way tables by medians'". *Understanding Robust and Exploratory Data Analysis, eds D. C. Hoaglin, F. Mosteller and J. W. Tukey. New York: John Wiley & Sons.* ISBN 0471384917. pp. 165–210. (1983)

[14] R. Gentleman, W. Huber, V. J. Carey, R. A. Irizarry, S. Dudoit, "Bioinformatics and Computational Biology Solution Using R and Bioconductor"

[15] I. H. Wittel, E. Frank, "Data Mining: Practical machine learning tools and techniques", *2nd edn. Morgan Kaufmann, San Fransisco* (2005)

[16] G. Stasis, A. C. Loukis, E. N. Pavlopoulos, S. A. Koutsouris, D. "Using decision tree algorithms as a basis for a heart sound diagnosis decision support system, Information Technology Application in Biomedicine". *4th International IEEE EMBS Special Topic Conference*, 2003

[17] Y. Yu, J. Khan, C. Khanna,L. Helman, P. S. Meltzer, G. Merlino, "Expression profiling identifies the cytoskeletal organizer ezrin and the developmental homeoprotein six- 1 as key metastatic regulators", *Nature Medicine* 10 (2004) 175–181.