Combining Singular Value Decomposition and t-test into Hybrid Approach for Significant Gene Extraction from Microarray Data

Mohammed Alshalalfa Reda Alhajj Jon Rokne

***Abstract*— Significant gene extraction from microarray data is a challenging problem which is of great interest to researchers in Computational Biology, Medicine, Computer Science and Statistics. A number of methods have been proposed for extracting the smallest number of genes which can accurately classify different samples. Most of these methods ignore the fact that microarray data is mostly noisy. For instance, only using a statistical t-test has been shown to be insufficient since it result in a high false discovery rate. Recently, a singular value decomposition (SVD) based approach was proposed for time series microarray data reduction, however it turned out not to be efficient for classifying microarray data. To overcome the shortcomings of these approaches, this paper proposes two methods to reduce false discovery rates. The first method involves an iterative t-test which finds the p-value for each gene under perturbation by eliminating one sample at a time. It eliminates weak noisy genes by dropping any gene which does not show significant p-value under all the conditions. The second method is a hybrid process which adapts a combination of the SVD and the t-test. It considers the entropy of all the data, and thus takes the correlation between genes into account. Classification accuracy is used to validate the significance of the extracted genes. The reported test results on two datasets demonstrate the applicability and effectiveness of the two proposed methods.**

Keywords: gene extraction, gene reduction, entropy, microarray data, t-test, singular value decomposition, p-value.

I. INTRODUCTION

The microarray technology is a powerful technique used to study the simultaneous expression of thousands of genes. This technique is mainly used to analyze gene expression of the genome under different conditions, such as time series cell cycles [7] and repeated samples in the tumor versus normal classification problem, e.g., [6]. Analyzing microarray data is significant for understanding the molecular mechanisms of the genes who control each other, and for understanding how genes behave under different conditions [7]. Data mining techniques such as clustering [8], classification [9], and association rules [10] have been successfully applied to microarray data in different contexts.

Clustering is used to group genes that have similar expression patterns under different conditions. The resulting clusters have been shown to have genes sharing similar functionality. The technique can be used to find the function of newly discovered genes by studying the functions of the genes in the same cluster. Clustering can also be used for sample class prediction [6]. Some of the most commonly used clustering algorithms include k-means [11], SOM [11], FCM [12], etc. While clustering is an unsupervised learning technique, classification is a supervised learning technique widely used to analyze gene expression data [6]. Classification requires some known classes, and requires having training and test data sets for building and testing a classification machine. First, the classification machine is trained with the training set, and then the accuracy of the machine is evaluated based on the test set. There are many classification techniques that have been used for microarray data analysis, including support vector machine (SVM) [9], [13], neural networks [18], [19] and k-nearest neighbor (k-NN), among others. Away from clustering and classification, association rule mining has not yet been extensively used for gene expression data analysis [14] since it is more difficult to apply and since it it difficult to interpret the results. For data mining techniques to be efficient and effective, feature reduction is important as preprocessing step.

Microarray data items have more than 10,000 gene values. Many of these data items represent genes which are not significant biologically and statistically. Such data items represent noisy genes that negatively affect clustering or classification. The aim of feature selection is to eliminate the data for genes which are not significant; for example genes which have many missing values, or genes that do not exhibit significant change between the samples. There are many benefits of feature reduction in biological data. First, feature reduction methods reduce the size of the data; hence, reduce computational cost. Second, the selected genes are very relevant to the experimental sample. Here, the objective of feature selection is to find the set of genes whose relevance to the experimental sample is maximal and the redundancy is minimal. Most of the proposed algorithms solve one of these problems. Integrating different approaches can therefore solve complimentary problems [15], [16].

There are a number of possible feature selection techniques. Each technique has specific assumptions for the feature selection. Statistical tests like unpaired t-test and F-test [17], [15], [16] are very good methods for feature selection. The disadvantage of these statistical tests is that a threshold value is required for selecting the top genes. Also, redundant genes can not be eliminated using these statistical tests since these methods do not take the complete data into account. They just evaluate the significance of each gene individually, and select the top ones, depending on the

M. Alshalalfa is with the Department of Computer Science, University of Calgary, Calgary, Alberta, Canada, msalshal@ucalgary.ca

R. Alhajj is with the Department of Computer Science, University of Calgary, Calgary, Alberta, Canada, he is also affiliated with the Department of Computer Science, Global University, Beirut, Lebanon, alhajj@ucalgary.ca

J. Rokne is with the Department of Computer Science, University of Calgary, Calgary, Alberta, Canada, rokne@ucalgary.ca

threshold. Other algorithms such as SVD [1] consider the complete data and assign a weight for each gene. Even if the data does not include genes which are significantly different among samples, SVD still returns the set of genes which have the highest entropy. So, the motivation for the work described in this paper was to investigate the applicability and analysis of these approaches for significant gene extraction from microarray data.

In this paper, we propose two approaches for gene selection where the main target is to reduce false discovery rates. The first approach is based on an iterative t-test for determining the p-value for each gene under perturbation by eliminating one sample at a time. In this manner, we eliminate weak noisy genes by neglecting all genes which do not show significant p-value under all conditions. The second method is a hybrid approach that combines the SVD and the t-test by considering the entropy of all the data which takes the correlation between genes into account. Classification accuracy is used to validate the significance of the extracted genes. The reported test results on two popular datasets, namely AML/ALL cancer data and breast cancer data, demonstrate the applicability and effectiveness of the two proposed approaches.

The remainder of this paper is organized as follows. Section II covers the related work. Section III presents the necessary background and describes the two proposed approaches. Section IV reports test results and the analysis. Section V presents a conclusions and some suggestions for future work.

II. Related Work

The group led by Golub [6] may be considered to be the first group who attempted to distinguish between two cancer types using gene expression data by considering the AML and ALL cancer subtypes. They used SOM classification model in combination with a weighted voting scheme for feature reduction. They obtained a strong prediction for 29/34 samples in the test data using 50 genes. Furey *et al* [20] applied SVMs to the AML/ALL data and derived significant genes based on a score calculated from the mean and standard deviation of each gene type. Tests were performed for 25, 250, 500, and 1000 top ranked genes. At least two test examples were misclassified in all the reported SVM tests.

Li and Wong [21] used a new feature selection method called emerging patterns. When they applied their method to the AML/ALL data, they were able to identify one gene (zyxin), which was able to classify 31/34 of the samples. Toure and Basu [19] applied a neural network methodology to cancer classification where 10 genes were used for classification purposes. Their neural network was able to fully separate the two classes AML/ALL during the training phase. However, the classification of the test set samples did not achieve high accuracy since 15 samples were misclassified. Zhang and Ke [13] applied SVM and CSVM for classification of the 50 genes reported by Gloub *et al*; two misclassifications occurred while using SVM, but no errors were reported when CSVM was used.

Entropy and perturbation based gene selection methods were also proposed for identifying significant genes from microarray data. Varshavsky*et al* [1] used SVD-entropy based ranking approach to select genes which change their expression along several samples. Their work aimed to reduce dimensionality of data for better clustering; however, they did not consider applying the same approach for extracting genes which can distinguish between two samples. In another paper by Varshavsky *et al* [22], they applied perturbations by eliminating up to 50% of the data to discover genes having similar expression profiles. Varshavsky *et al* did not consider the weight of the samples when eliminating the samples. Also, since the number of samples in the microarray data is usually small, eliminating 50% of the samples may lead to information loss.

III. The Applied Methodology and Approaches

This section first introduces the basic ideas of unpaired t-test and singular value. These techniques form the basis for the two approaches which are then proposed for gene selection,

A. T-test

The unpaired t-test is a statistical test applied to data containing two or more groups. The test assesses whether the means of two groups are statistically different from each other. The null hypothesis is in this case that *the means of each gene in the two samples are equal*, i.e , $H_0 : \mu_1 = \mu_2$. Given the replicas of particular treatment and control samples, it is possible to compute the t-test for any gene g for differential expression by using the following formula under the assumption that genes have differing standard deviations [5]:

$$t_g = \frac{\overline{x}_{g,t} - \overline{x}_{g,c}}{\sqrt{\frac{s^2_{g,t}}{n_t} + \frac{s^2_{g,c}}{n_c}}}. \quad (1)$$

Here $\overline{x}_{g,t}$ and $\overline{x}_{g,c}$ are the means of replicas of treatment and control conditions with respective standard deviations $s^2_{g,t}$ and $s^2_{g,c}$, and replica counts n_t and n_c for gene g. It is clear that t-test favors samples with large mean differences and small standard deviations. Statistically, when the null hypothesis is rejected, there is a probability of wrong rejection, i.e., the decision is not 100% correct. This probability of being uncertain about the decision is expressed as the *p-value*. The p-value is therefore an important measure for the uncertainty of a particular decision, e.g., gene g is differentially expressed.

B. SVD based Gene Selection

Given an $M \times N$ matrix A, the singular value decomposition (SVD) of A is its representation as $A = UWV^T$, where U is an orthogonal $M \times M$ matrix; V is an orthogonal $N \times N$ matrix; and for the diagonal matrix W, elements are non-negative numbers in descending order. The singular value decomposition has many useful properties. For instance, it can be used to solve underdetermined and overdetermined systems of linear equations, find inverse and the pseudo-inverse matrices, compute the matrix condition number and

calculate the vector system orthogonality and orthogonal complement. SVD has several applications in areas such as signal processing, information retrieval [3], and recently gene expression data analysis, e.g., [1], [4], [2]. SVD can be applied to the problem of grouping genes by transcriptional response, and grouping assays by expression profiles. SVD also helps in the search for biologically meaningful signals in noisy data.

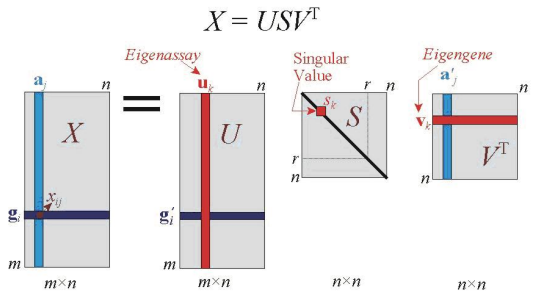

Fig. 1. SVD based Microarray Analysis (adapted from [2])

Most of the existing feature filtering methods do not consider the complete data, thus the filtered genes do not represent the information in the original data. SVD is a method that avoid such problems since it can reduce the size of the data into a smaller number of features without losing any knowledge from the original data. Varshavsky *et al* [1] have demonstrated the effectiveness and applicability of SVD for feature extraction. Figure 1 shows a microarray data of m genes and n sample. Using SVD, it is possible to extract eigengenes which represent the original data.

SVD is a linear transformation of the expression data from n-genes by m-array represented by a matrix $A_{m\times n}$ to the reduced diagonal L-eigengenes by L-eigenarrays matrix, where $L = \min(n, m)$ [4] and $s_i, i =, \ldots L$ are the singular values. Alter *et al* [4] calculated the normalized relative significance p_k of the k-th eigengene for $A_{m\times n}$ as follows:

$$p_k = \frac{s_k^2}{\sum_{i=1}^{L} s_i^2} \tag{2}$$

and the Shannon entropy of the data represented by $A_{m\times n}$ is calculated as:

$$E(A_{m\times n}) = -\frac{1}{log(L)} \sum_{k=1}^{L} p_k log(p_k). \tag{3}$$

Varshavsky *et al* [1] have defined the contribution of the i-th gene CE_i by a leaving out comparison as:

$$CE_i = E(A_{m\times n}) - E(A^{(i)}_{m-1\times n}) \tag{4}$$

where $A^{(i)}_{m-1\times n}$ is the matrix $A_{m\times n}$ with the i-th row deleted. The SVD-based approaches discover genes which show change in expression levels across samples. However, there has been no attempts to apply SVD for selecting differentially expressed genes across two samples for classification purposes. One of the contributions in this paper is therefore the adaptation of SVD to select significant genes across samples as explained in Section III-D.2.

C. Classification and Support Vector Machine

Classification is a supervised technique that categorizes a given set of instances into classes of know behavior. The aim of classification may be stated as follows: *to build a model which establishes a set of classes from a given training set which can determine the most appropriate classes in this set to which new training set compatible data points belong*. The more comprehensive the available training data set is, the more the technique learns and the more accurate results are produced.

The Support Vector Machine (SVM) is one of the most powerful classification techniques [24]. The SVM tries to find a hyperplane between two classes and it maximizes the distance between the points and the hyperplane. For complex data, the points are transformed into high dimensional feature space and the transformation may be non-linear, for example, polynomial, Radial basis function or sigmoid. The aim of the transformation is to make it possible to define a hyperplane in the high-dimensional feature space which can separate the classes. SVM has shown to be efficient and accurate classification technique for microarray data [13], in addition to being efficient in significant point extraction. SVM is easy to understand and it is easy to interpret the results, but the implementation is difficult as the mathematics behind SVM is complex and require extra effort to understand.

D. The Proposed Approaches

We now describe the two approaches proposed for gene reduction where the first approach is based on an iterative t-test while the second approach mainly integrates SVD into the process.

1) Iterative t-test based approach: A requirement for microarray data preprocessing is removal of noisy genes. The result of a number of systematic errors both in the microarray and the image processing steps is that some genes show a very high expression level under one sample in a class while the other genes in the same class show a low expression level. These kinds of outlier expression levels should not affect the gene selection process. Using the regular t-test does not eliminate such outliers. Consequently, we have tested how one sample elimination can affect the gene selection process using the t-test. We applied the t-test to the genes to extract those which show significant patterns under perturbation. We then apply perturbations by removing samples one by one and find the p-values for the genes under all conditions. We eliminate one gene at a time in order to avoid information loss. When we remove one sample, we find all genes whose p-values are less than a threshold which is set to 0.001 for example. Then we generate a matrix called $Significant\ Genes$, where each row contains significant genes under certain condition (removal of one sample) and from this we find the most significant association rules by considering the frequent set(s) with the maximum support value. Surprisingly the tests reported rules with 100% support.

We sort the frequent sets in descending order by their support value, and then we consider the genes that appear in

the rules that have the highest rank. The process applied in this study can be summarized as follows: If the gene has a p-value less than the threshold under all the conditions, then it is significant. After getting the significant genes, they are then processed using SVM for classification. The results have been compared with regular t-test and we have shown that the genes eliminated by our approach can be considered to be false positives since they have low classification accuracy. To summarize this approach, we eliminate the first sample in the data and we find the p-value for each gene using t-test. Genes with p-value less than the threshold are stored in the first row of $Significant\ Genes$ matrix. Then, we return the first sample and eliminate the second; we find p-values for all genes and store the genes whose p-values are less than the threshold in the second row of the matrix. We repeat the same process for all samples, i.e., at step $i > 1$, we return sample $(i-1)$ and remove sample i. Finally, we take the genes found in every row as significant and not false positives.

2) SVDttest Approach: The SVD-based approach proposed by Varshavsky *et al* [1] is not appropriate for gene extraction from multi-class data. The reason for this is the following. Assume we have a microarray data set with two classes each having two samples denoted by $Class1_S1$, $Class1_S2$, $Class2_S1$, $Class2_S2$. A gene having data values [0,0,1,1] should be significant for distinguishing between the two classes. However, the SVD approach by Varshavsky *et al* considers a gene whose values are [0,1,1,0] as significant, although it should not be. This led us to adapt the SVD approach to two class data in order to extract significant genes. The importance of each gene is computed as in Equation 4. Genes with high E value are selected as important. In order to adapt the SVD approach to the binary classification problem, we need to compute the average for the values of each gene under each class. In this manner, the dimensionality of the data is reduced from $m \times n$ to $m \times 2$. This reduction helps us to identify genes which have high entropy due to sample difference. The SVD-based approach considers the entropy of the gene with respect to the other genes in the data and t-test considers the data distribution for each gene. Combining both the SVD and the t-test will provide a better indication of significance of each gene. To implement this combination, we have defined a new term, denoted $SVDttest$, which is computed as the ratio of SVD over t-test:

$$SVDttest(g) = \frac{CE_g}{t_g} \tag{5}$$

where CE_g is computed by Equation 4 and t_g is computed by Equation 1. Based on extensive testing and analysis of the results, we realized that genes with $SVDttest$ value greater than 1 may be considered to be significant.

The algorithm proposed assumes that a full microarray data set is given with the property that the data set has two classes each having many samples.

We then reduce the dimensionality to two by averaging the samples in each class. We use Equation 4 to calculate the entropy of each gene which shows how the entropy of the matrix is affected when the gene is removed. If entropy does not change then this indicates that the gene is not important. The significance of the gene increases, however, as the change in the entropy increases. The advantage of reducing the dimensionality of the genes to two means that difference across samples indicate genes with high entropy. The method proposed by Varshavsky *et al* does not ensure that high entropy genes are due to the difference in classes. It just ensures that the genes have dynamic gene expression profiles along the samples, but not necessarily across classes. The new approach proposed in this paper considers both statistical and entropy based significance for each gene. In summary:

1) Find the p-value of each gene using t-test
2) For each gene, average the gene expression value under each condition, i.e., if there are two classes of data then the result is an $N \times 2$ data matrix, where N is the number of genes.
3) Find the contribution of each gene to the entropy of the matrix using SVD as in Equation 4.
4) For each gene, divide the entropy contribution calculated in step 3 by the p-value from step 1, and select genes with $SVDttest$ value greater than 1. The tests conducted demonstrate that the larger the score, the more significant the gene is.

For a gene to have high a $SVDttest$ value, it has to have either a very large SVD value due to the difference across classes, or a very small p-value due to large difference across classes.

IV. Experiments and Analysis

Data preprocessing and the experiments were conducted using matlab. Gene selection was performed using the *t2test* function implemented in matlab. The LIBSVM package, a free library for classification and regression implemented in matlab, was used for classification. The code for the SVD based gene selection was provided by Varshavsky *et al* [1]. We have run the programs on an Intel machine with Core2Duo CPU 2.0GHz and 1.99GB of RAM running Windows XP professional version 2002 SP2. For biological analysis, we used the STRING database (http://STRING.embl.de/). The experiments were conducted on two data sets: namely AML/ALL and Breast cancer; both of which are described below.

TABLE I
CLASSIFICATION RESULTS FOR AML/ALL DATA USING 40 GENES FILTERED BY T-TEST

	Linear SVM	**Polynomial SVM**	**RBF SVM**
Accuracy	*94%*	*91 %*	*97 %*
Cross-validation	*100%*	*97%*	*97%*

TABLE II
CLASSIFICATION RESULTS FOR AML/ALL DATA USING 25 GENES FILTERED BY ITERATIVE T-TEST

	Linear SVM	**Polynomial SVM**	**RBF SVM**
Accuracy	*94%*	*94 %*	*94 %*
Cross-validation	*100%*	*100%*	*100%*

TABLE III

CLASSIFICATION RESULTS FOR AML/ALL DATA USING 13 GENES FILTERED BY SVD-TTEST

	Linear SVM	Polynomial SVM	RBF SVM
Accuracy	*97%*	*97 %*	*97 %*
Cross-validation	*92%*	*92%*	*97%*

TABLE IV

CLASSIFICATION RESULTS OF THE 15 ELIMINATED FROM AML/ALL DATA USING ITERATIVE T-TEST

	Linear SVM	Polynomial SVM	RBF SVM
Accuracy	*58%*	*61%*	*58%*
Cross-validation	*100%*	*97%*	*100%*

TABLE V

COMPARISON AMONG T-TEST, SVD, AND SVD-TTEST CUTOFF VALUES

Gene index	SVD-ttest	t-test	SVD
3320	*4.3974e+005*	*1.1077e-010*	*4.87e-05*
2121	*347.5*	*1.9935e-007*	*6.93e-05*
6806	*152.33*	*7.813e-007*	*11.9-05*
3258	*101.03*	*7.1392e-007*	*7.21e-05*
804	*13.642*	*4.0577e-006*	*5.54e-05*
2111	*13.18*	*3.2612e-006*	*4.30e-05*
2186	*10.737*	*9.9251e-006*	*10.7e-05*
5501	*10.184*	*6.3121e-006*	*6.43e-05*
4328	*9.4266*	*8.6175e-006*	*8.12e-05*
1928	*9.3483*	*3.5036e-006*	*3.28e-05*
4211	*4.0965*	*1.0748e-005*	*4.40e-05*
1673	*2.9822*	*1.0747e-005*	*3.21e-05*
1704	*2.2168*	*3.8127e-005*	*8.45e-05*

A. AML/ALL Data

The AML/ALL data was obtained from Golub *et al*. This data contains 7130 genes for a sample of 73 patients, where 38 samples are for training of which 27 are AML and 11 ALL, and 35 for testing of which 23 are AML and 12 ALL. We have selected genes which have at most 8 missing values. A missing value means that the spot was not identified. As a part of the preprocessing step, the missing values were predicted according to the nearest neighbor values and the data was log transformed. We first filtered the data using unpaired t-test with p-value=0.001. As a result, 40 genes were selected and passed to SVM. The achieved accuracy is shown in Table I. The iterative t-test was then applied to the same set of genes. The 40 genes were filtered down to 25 genes, which showed to be significant at each perturbation condition. The SVM classification results using the 25 genes are shown in Table II. We also derived the classification results of the 15 genes filtered out as shown in Table IV. Afterwards, we applied the proposed SVD-ttest method on the same set of genes. We even reduced the number of genes to 13 and got better accuracy as shown in Table III. The only sample misclassified was sample 70. This has been reported in the literature as the most difficult sample to be correctly classified along with the two samples 66 and 67 [23]. Here, it is worth mentioning that we also applied SVD alone for filtering, but then the results obtained were very poor. We have highlighted the advantage of $SVDttest$ in Table V, where the reported results demonstrate how it is easy to make a cutoff value using the proposed approach. Also, it is important to note that the order of the genes in the proposed approach is different in t-test and SVD. We also studied the biological functionality of the selected genes

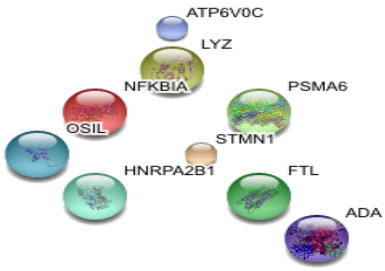

Fig. 2. Protein interactions network of the genes selected from AML/ALL

using the STRING database. The results showed that the proteins of the selected genes do not interact with each other. This indicates that functionality redundancy was reduced. The protein interaction network for the selected genes is shown in Figure 2.

Using AML/ALL data, we have shown that the iterative t-test successfully eliminated false positives. The eliminated genes showed poor accuracy as reported in Table IV. We also, showed that the genes selected by our approach are efficient biomarkers. In Table V, we summarized the rank of genes in different methods. We highlighted the fact that the order of the top ranked genes in each method is different and that the top genes are not the same in the two approaches. We also illustrated that it is easier to decide on the cutoff value using our method rather than solely applying the t-test.

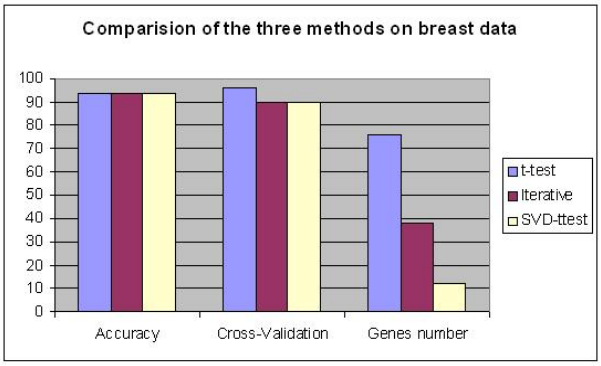

Fig. 3. Comparison among the three methods on breast data using RBF

B. Breast Data

Breast cancer data was obtained from [25]. It has 7129 genes and 47 samples. The samples were 23 estrogen receptor positive split as 15 for training and 8 for testing, and 24 estrogen receptor negative split as 15 for training and 9 for testing. The data was log transformed. We used the same three methods for gene filtering and the three SVMs for classification. We reported the comparison of the three methods in Figure 3, namely the t-test, the iterative t-test and the $SVDttest$, using RBF as a classifier. The results reported showed that the small number of genes filtered by the proposed approach has the same classification accuracy as the genes filtered by the other approaches. Our main aim for this data is to show that our method requires less number

of genes while obtaining almost the same accuracy. We also applied the iterative approach and the biological analysis on the selected genes, but the results are not reported in this paper due to space limitations.

C. Discussion

As part of microarray data preprocessing, significant gene selection is crucial for better and accurate classification. T-test has been widely used for gene selection, but choosing the threshold is very critical and has absolute boundary; if we set the threshold to be 0.01, then we will select a gene which has p-value of 0.009999 and exclude a gene which has p-value as 0.0101. Furthermore, we may select all the genes in the data if all of them demonstrate to be statistically significant among samples. The reason for this is the lack of the ability to consider the whole data while selecting the genes. On the other hand, the SVD based approach proposed by Varshavsky *et al* [1] does consider the whole data while selecting the genes; however, it still selects a set of genes even if no gene is statistically significant. The idea of SVD-ttest has been inspired by the limitations of those two methods. Proposing a method which can consider the statistical significant of the individual genes and their entropy on the whole data is very important. Another advantage of the proposed approach is that there is no need for a cutoff value. Statistically significant genes with large entropy are selected. There still does not exist a solid interpretation supporting this, but experimentally it showed to be working very well. Analyzing the biological importance of the selected genes, we have seen that they participate in variant processes in the cell like cellular iron ion homeostasis, cell differentiation, proteolysis and T-cell activation and cell-cell adhesion. In addition, the genes selected from AML/ALL do have role in apoptosis, leukotriene biosynthesis, ubiquitine -dependent protein catabolic process, and inflammatory responses in addition to cytoskeletal anchoring. This shows that the selected genes do not have common functionality among them and they do represent most of the cellular functionalities related to cancer cells. Interestingly, we have seen that two of the selected genes are involved in iron transport, which makes the iron transport process a target for more investigation about the exact role of iron in AML or ALL.

V. Conclusions and Future Work

In this work, we have proposed two methods to extract significant genes from classification microarray data. The first approach considers the noisy data and eliminates genes which are noisy. The second eliminates the genes which do not show high entropy and statistical significance. The tests conducted demonstrate the significance of the proposed approaches as interesting contributions for more appropriate gene selection. As a result, the proposed approaches significantly reduced false discovery rate. After we have experimentally demonstrated the power of the proposed approaches, we are currently concentrating on developing a stronger mathematical model which combines both t-test and SVD differently from the direct ratio. Our target is a more robust approach.

References

[1] R. Varshavsky, A. Gottlieb, M. Linial and D. Horn, Novel usupervised feature filtering of biological data, *Bioinformatics*, vol. 22, 2006, pp.507-513.

[2] M. Wall, A. Rechtsteiner and L. Rocha, Singular value decomposition and principal component analysis,*A practical approach to microarray data analysis*, 2003.

[3] M. Berry, Z. Drmac and E. Jessup, Matrices, Vector Spaces, and Information Retrieval, *SIAM Review*, vol.41, 1999, pp.335-362.

[4] O. Alter, P. Brown and D. Botstein, Singular Value decomposition for genome-wide expression data processing and modeling, *PNAS*, vol.97,2000, pp 10101-10106.

[5] S. Dudoit, Y.H. Yang, M. Callow and T. Speed, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Technical Report #578, University of California, Berkeley*, 2000.

[6] T.R. Golub, et al, Molecular Classification of Cancer: class discovery and class prediction by gene expression monitoring, *Science*, vol.286, 1999, pp.531-537.

[7] N. Banerjee and M. Zhang, Identifying cooperative among transcription factors controlling cell cycle in yeast, *Nucleic Acids Research*, vol.31 , 2003, pp.7024-7031.

[8] I. Belitskaya, A generalized clustering problem with application to DNA microarray, *Statistical Applications in Genetics and Molecular Biology*, vol.5, 2007.

[9] K. Kianmehr and R. Alhajj, Support vector machine approach for fast classification, *DaWaK*, 2006.

[10] F.J. Lopez, A. Blanco, F.Garcia, C. Cano and A.Marin, Fuzzy association rules for biological data analysis: A case study on yeast, *BMC Bioinformatics*, vol.9, 2008.

[11] M. Alshalalfa and R. Alhajj, Application of double clustering to gene expression data for class prediction, *AINA*, 2007.

[12] J. Wang, T.H. Bó, I. Jonassen, O. Myklebost and E. Hovig, Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data,*BMC bioinformatics*, 2003.

[13] Z. Xuegong and K. Haixin, AML/ALL cancer classification by gene expression data using SVM and CSVM approach, *Genome Informatics*, 2000, pp.237-239.

[14] M. Kabbaz, K. Kianmehr, M. Alshalalfa and R. Alhajj, Fuzzy classifier based feature reduction for better gene selection, *LNCS*, 2007.

[15] J.J. Chen, C-A. Tsai, S. Tzeng and C-H. Chen, Gene Selection with multiple ordering criteria, *BMC Bioinformatics*, 2007.

[16] A.O. Hero, Gene selection and ranking with microarray data,*Signal processing and its applications*, 2003.

[17] J.J. Chen, S-J. Wang, C-A. Tsai and C-J. Lin, Selection of differentially expressed genes in microarray data analysis, *The Pharmacogenomics Journal*, vol.7, 2007, pp.212-220.

[18] S. Bicciato, M. Pandin, G. Didon and C. Di Bello, Pattern identification and classification in gene expression data using an autoassociative neural network model, *Biotechnology and Bioengineering*, vol.81, 2002, pp.594-606.

[19] A. Toura and M. Basu, Application of neural network to gene expression data for cancer classification, *Proc. of IEEE International Joint Conference on Neural Networks*, 2001, pp.583-587.

[20] T.S. Furey, et al, Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, vol.16, 2000, pp.906-914.

[21] L. Jinyan and L. Wong, Identifying good diagnosis gene group from gene expression profile using the concept of emerging patterns, *Bioinformatics*, vol. 18, 2002, pp 725-734

[22] R. Varshavsky, A. Gottlieb, D. Horn and M. Linial, Unsupervised feature selection under perturbations: meeting the challenges of biological data,*bioinformatics*, vol.23, 2007, pp.3343-3349.

[23] S. Dudoit, J. Fridly and T.P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data. *J.Am.Stat.Assoc.*, 1997, pp.77-87.

[24] M. Pirooznia, J. Yang, M. Yang and Y. Deng, A comparative study of different machine learning methods on microarray gene expression data, *BMC Genomics*, 2007.

[25] B. West, et al, Predicting the clinical status of human breast cancer by using gene expression profiles, *PNAS*, 2001, vol.98, pp.11562-11567.