# A New Gene Selection Technique Using Feature Selection Methodology

**Noushin Ghaffari**
**University of Houston – Clear Lake**
**Houston, Texas 77058**
**Ghaffarin8978@UHCL.EDU**

**Hisham Al-Mubaid**
**University of Houston –Clear Lake**
**Houston, Texas 77058**
**hisham@UHCL.EDU**

## Abstract

The DNA Microarray technology can measure the expression levels of thousands of genes simultaneously, and produces huge volumes of gene expression data. Such gene data include complex variations among expression levels of genes in the various classes of samples, which allows for classifying and clustering the samples based on only a small subset of genes. We aim to identify those genes that demonstrate high capabilities of discrimination between the classes of samples (*e.g. the normal vs disease tissue samples*). We present a new technique for gene selection and extraction using various feature selection techniques. Our method is based on computing thresholds and discriminating capabilities of each gene, and classifying the data according to only those genes that have highest discriminating capabilities. The method extracts very small subsets of *informative* genes that can improve the classification accuracy. We applied the method on four different common gene expression datasets used mainly for this purpose. The method produces encouraging and competitive results of classification performance compared with recent similar techniques.

## 1. INTRODUCTION

DNA Microarray chips were first introduced in 1995 for measuring the expression levels of thousands of genes simultaneously [1, 2, 13]. This technology studies the genes with known sequence. These genes are amplified by Polymerase Chain Reaction (PCR) technique. A robot spots the PCR resulted genes onto an ordinary glass microscope slide. The next process denatures and links the spotted DNA to the glass side [13]. Each microscope slide contains a grid-like pattern like an array with thousands of spots of amplified copies of each gene. Immobilized DNA on the microarray will be hybridized with a probe, which is a known labeled DNA sequence. In order to make the probe, mRNA is isolated from control or diseased samples, and converted into cDNA. The nucleotides used to produce cDNA include a green dye called Cy3, or a red dye called Cy5. Therefore, cDNA is distinguishable by colors. After passing through the entire process, sensors and scanners are used to detect dyes (green and red) and record their location and intensities into the computerized system [13, 7]. Since each microarray contains thousands of DNA spots, the output numeric data is too much to be processed manually. Therefore, there is a great demand in the biomedical field for efficient methods for analysis and manipulation of gene expression data. These data includes complex variations among expression levels of each gene in the normal vs disease tissue samples, which allows for classifying and clustering the samples into *normal vs disease* based on only a small subset of the genes. For example, one experiment carried out on samples of lung cancer tissue and samples of normal tissues can produce the expression results of thousand or maybe tens of thousands of genes for the normal and disease tissue samples. Each one of these genes may have variations in its expression levels between the *normal* versus the *disease* samples. We should mention that classes of samples can also be types of some disease (*e.g. cancer*) such that each class represents a different type of the disease.

The goal of this work is to extract those genes that demonstrate high discriminating capabilities between the normal and disease samples. We propose a new method for gene classification and extraction using various feature selection techniques. Our method is based on computing thresholds and discriminating capabilities of each gene and classifying the data according to only those genes that have highest capabilities to discriminate between the two classes (viz. *normal, disease)* of samples. The method extracts very small subsets of useful salient genes that can improve the classification accuracy of tissue samples. We applied the method on four different gene expression datasets used commonly for this purpose. Our proposed method produces encouraging and competitive results in terms of classification performance compared with recent similar techniques.

Related work: During the last few decades, a number of methods and algorithms have been proposed and applied into gene expression profiles; some of them produced very significant results in terms of accuracy. Paul and Iba (2005) [1] modified the Probabilistic Model Building Genetic Algorithm (PMBGA) into a Random PMBGA (RPMBGA) for gene selection and applied it to three datasets. They have tried to reduce the size of gene subsets while keeping accuracy of classification in the high level. In all three datasets they have results superior or comparable to previous researches. For the same task, Liu et al. (2004) [2] used the neural network for gene expression profiles. They used 100 iterations of resampled data as an input to their architecture, which consists of three neural network feature selection methods: *Ranksum* test, Principle Components Analysis (PCA) and clustering. They used Kent Ridge datasets [17] and found 100% accuracy for ALL-AML and Lung Cancer datasets, and 97.06% for the prostate cancer dataset [2]. In another effort, Gordon et al. (2002) [3] tried to distinguish between the pathological distinction of malignant mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung cancer using gene expression levels only. Their focus was on searching all of the genes represented in the microarray for genes with a highly significant difference in average expression levels between the two types MPM and ADCA. They chose 8 top genes. Five of the genes correctly classified into MPM and 3 into ADCA. They validated their results using an independent testing set [3].

Shen et al. [5] used dimension reduction for cancer classifications. They combined *Penalized Logistic Regression* (PLR) techniques with *Partial Least Squares* (PLS) and with *Singular Value Decomposition* (SVD) separately. They found out that the combination of PLR and PLS is more preferable in terms of accuracy and computational speed. Furthermore, Y. Lee et al. (1999) [6] used Multi-category Support Vector Machine (MSVM) versus Binary SVM for cancer classification. They used leaving-out-one cross-validation (LOOCV) and generalized approximate cross-validation (GACV) for validation. Their result for *leukemia* data using MSVM resulted in a 0 to 1 test error, at best, which is a very good result. They applied their method into Small Round Blue Cell Tumors (SRBCT), as well [6].

The main objective of these methods is to select the most useful gene subset by applying dimensionality reduction into the gene expression profile. In spite of all, Do et al. (2003) [4] did not change the number of genes (features); instead they had improved Proximal Support Vector Machines (PSVM) method. They proposed a new column-incremental linear PSVM to deal with the huge amount of data by avoiding loading of the whole dataset in the main memory. Their reported accuracy for ALLAML Leukemia is 97.06%, prostate cancer is 97.06%, and is 98.66% for lung cancer.

## 2. THE PROPOSED METHODS

The representation of genes expression levels generated by the Microarray technology is a two-dimensional representation including two or more classes of (*tissue*) samples as columns and the genes as rows. We transpose the matrix and thus each column will be representing a gene and the samples will be the rows. Such a gene expression matrix contains large number of genes (usually *~20,000-25,000*) such that each gene has expression levels in the samples of "*class-1*", "*class-2*", "*class-3",…etc.* with a modest number samples (rows) in each class. In the simplest case there can be only two classes; for example, *class-1* representing *normal* tissue samples, and *class-2* representing *cancer* tissue samples. However, sometimes a dataset may include several classes, for example, each class represents one type of cancer. In this research we cast the multi-class problem as a two-class case and we deal with two classes at a time. The variation in the expression levels of each gene between *class-1* vs. *class-2* samples determines how much that gene is related to one of the two classes. We would like to determine how much a given gene discriminates between *class-1* and *class-2* samples. The gene that demonstrates high differences in its expression levels between *class-1* and *class-2* is a good "*informative"* gene that is typically highly related with the disease of *class-2* samples (assuming *class-2* is disease tissue samples). Such informative genes can produce high accuracy in the process of samples classification. We want to identify the highly informative genes to be used for classification of samples instead of using the entire set of genes. Our method is based on computing a discriminating value (*v*) for each gene in the dataset. A gene with the highest "*v value*" is the one that have the highest differences in its expression levels between the two classes. Then we sort the genes based on their computed *v* values. And then, we select the top *n* genes and delete the remaining (unselected) genes from the data. Then we use machine learning determine how accurately that small subset of *n* genes can classify the samples.

### Computing *v* values

Suppose we are given a gene expression matrix with two classes of samples *classe-1* and *class-2*, and each gene is represented as a column. Assumer, further, that we have a threshold value *t*. For each gene we define four values *a*, *b*, *c*, and *d* as follows:

$a$ = # of gene expression values of gene $g$ in *class-1* $\geq$ t

$b$ = # of gene expression values of gene $g$ in *class-1* $<$ t

$c$ = # of gene expression values of gene $g$ in *class-2* $\geq$t

$d$ = # of gene expression values of gene $g$ in *class-2* $< t$

For example, if the threshold $t = 0$, then we compute for a given gene how many of its expression values in *class-1* are above or equal to 0 (*a* value), or below 0 (*b* value); and how many of its expression values in *class-2* are above 0 (*c* value), or below 0 (*d* value). If a gene *g* has its four values (*a,b,c,d*) as follows $a = |class\text{-}1|$, $b = 0$, $c = 0$, and $d = |class\text{-}2|$. That is, all its values in *class-1* are above the threshold, and all its values in *class-2* are below the threshold. In other words, there is a threshold, *t*, that clearly separates the gene values in *class-1* from its values in *class-2*. Then we say that this gene, *g*, perfectly differentiates between *class-1* and *class-2* in its expression levels, and it's a very useful gene. Furthermore, for a given threshold *t* the most useful gene is the one that has the highest *a* and *d* values and lowest *b* and *c* values.

| Dataset | Number of Genes | Number of Classes | Training Set | Testing Set |
|---|---|---|---|---|
| ALL-AML Leukemia [6] | 7129 | 2<br>AML vs. ALL | 38 samples: 27 ALL and 11 AML | 34 samples: 20 ALL and 14 AML |
| Lung cancer [3] | 12533 | 2<br>MPM vs. ADCA | 32 samples: 16 MPM and 16 ADCA | 149 samples:<br>134 ADCA and 15 MPM |
| Prostate Cancer [2] | 12600 | 2<br>tumor vs. normal | 102 samples: 52 tumor and 50 normal | 34 samples: 25 tumor and 9 normal |
| DLBCL [7] | 4026 | 2 germinal vs. activated | 47 samples: 24 germinal and 23 activated | NA |

**Table 1:** The four gene expression datasets used in our experiments

Then, the measure [ *( a + d ) – (b + c )* ] is a good indicator of how much a gene differentiates between the two classes. Thus, we compute for each gene a *v* score as follows:

$$v = ( a + d ) - ( b + c ) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

and we select the genes that have the highest *v* values. Next, we sort all genes (columns) based on their *v* values in descending order. Recall that the computed *v* values for all genes depend on the thresholds. Next, we discuss how we select thresholds.

**Selecting thresholds for computing v values:**

We want to find a threshold *t* that separates the gene expression levels in *class-1* for its levels in *class-2* according to *Equation (1).* We initially test the threshold t = 0, then we compute the *v* value for each gene based on *t=0*, then we sort them and select the *n* genes with the *n* highest *v* values. In our method, after careful and extensive experimentation we identified three techniques for selecting the threshold:

(1) The basic technique: t = 0.
(2) Use a separate threshold for each gene.
(3) Use the same threshold *t* for all genes where *t* can be selected from the set T={-1024, -512, -256, …, 2, 0, 2, 4, 8, ……., 512, 1024}.

We found that selecting the threshold from T, using method (3), gives better performance in classification accuracy most of the time.

**Learning and classification**

We have discussed how we select a subset of the *n* most significant genes from the entire set of genes. We remove all unselected genes and then the data has *n* columns. We want to evaluate the selected gene subset using machine learning. We do a *two-class* classification based on only the *n* selected genes. We use support vector machines (SVM) [15, 16] for learning and classification. SVM is an inductive learning technique for two-class classification. Numerous theoretical and empirical justifications exist in the literature to support SVM [16]. In our method, we take two classes at a time and apply SVM to train on these two classes and produce a classifier (*model*). The classifier will then be used in the classification phase to classify testing samples. We use the SVM-*light* implementation (svmlight.joachims.org*)* with the default parameters.

## 3. EXPERIMENTS AND RESULTS

**Datasets**

We used four different microarray datasets to evaluate our method: ALL-AML Leukemia [6], Lung cancer [3], Prostate cancer [2], and the Diffuse Large B-cell Lymphoma (DLBCL)[7]. Table 1 contains the details of these datasets (*most of these datasets were obtained from [17]*). These datasets are used commonly for sample classification and gene clustering research. For the DLBCL dataset the whole available microarray data is divided into training set and testing set.

**Experimental Setups:** *Data Preprocessing*

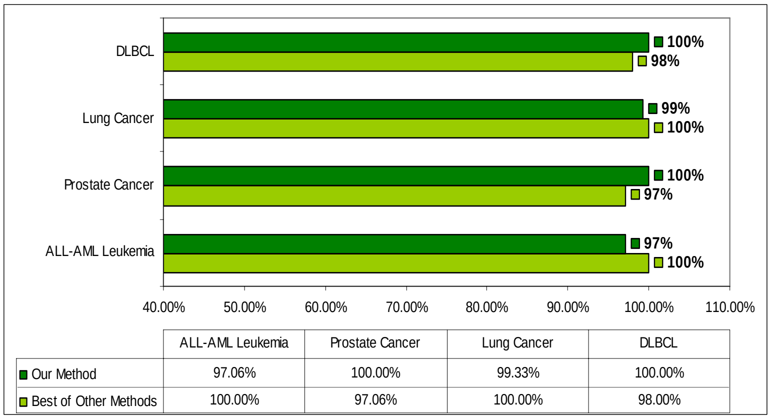
The four datasets were preprocessed from the raw Microarray data. Each ratio value in raw gene expression data is associated with a character: A for *active*, P for *hyperactive* and M for *silent* (normal). These characters are irrelevant in such tasks like ours; therefore all of these characters in all datasets were removed [17]. Also any other textual information about Affymatrix, Descriptions, Accessions or etc., added by producers of the raw data, were removed.

**Results and Discussion**

We ran our method to select subsets of most significant (useful) genes for a number of subset sizes. Subsets with size of 1, 2,… ,10, 50 and 100 genes were selected. The DLBCL dataset is the only available gene expression profile data contains 4026 genes. This dataset is divided into two individual sets, training and testing sets. For division 80/20 rule is used. The testing subset is selected from 3 different portions of data, beginning, middle and end.

| Number of Genes | Datasets | | | |
|---|---|---|---|---|
| | AMLALL<br>*Threshold=64* | Prostate Cancer<br>*Threshold=64* | Lung Cancer<br>*Threshold=512* | DLBCL<br>*Threshold=0* |
| 10 | 97.06% | 100.00% | 98.66% | 100.00% |
| 9 | 97.06% | 100.00% | 98.66% | 100.00% |
| 8 | 97.06% | 100.00% | 98.66% | 100.00% |
| 7 | 97.06% | 100.00% | 98.66% | 100.00% |
| 6 | **97.06%** | 100.00% | 98.66% | 100.00% |
| 5 | 91.18% | 100.00% | 98.66% | 100.00% |
| 4 | 94.12% | 100.00% | **99.33%** | 90.00% |
| 3 | 94.12% | 100.00% | 96.64% | **100.00%** |
| 2 | 88.24% | **100.00%** | 98.66% | 80.00% |
| 1 | 94.12% | 97.06% | 98.66% | 70.00% |
| **Average** | **94.71%** | **99.71%** | **98.53%** | **94.00%** |

**Table 2:** Results of classification accuracy for each dataset with its best corresponding threshold. 10 experiments with subset sizes from 1 to 10 genes are run for each dataset.



| | ALL-AML Leukemia | Prostate Cancer | Lung Cancer | DLBCL |
|---|---|---|---|---|
| ■ Our Method | 97.06% | 100.00% | 99.33% | 100.00% |
| □ Best of Other Methods | 100.00% | 97.06% | 100.00% | 98.00% |

**Figure 1:** Comparison between our results and other published results on the four datasets, the best results known up-to-date are taken from [2].

*Threshold: t = 0:* The basic threshold, *t=0*, assumes that all values in *class-1* are above zero and values in *class-2* are all negative. In this case definition of class positive and negative can impact the result tremendously. Each dataset is used for population sizes 1 to10 genes and t= 0. The average accuracy in each dataset is as follows:

| ALL-AML | Lung cancer | Prostate cancer | DLBCL |
|---|---|---|---|
| 89.77% | 86.18% | 95.90% | 94.0% |

*Using separate threshold for each gene:* In the second set of experiments we tested our method using the technique of separate threshold for each gene. In this case, the algorithm starts from first gene and passes through all genes. For each gene, the error rate (*b+c*) is calculated based on different numbers (*from lowest gene expression level to the highest gene expression value*). The number that produces the lowest error rate (b+c), is selected as a threshold, so each gene has its own threshold. In the next step program calculates the *v* score for each gene based on the calculated threshold. The genes with the highest *v* values will be selected. The average accuracy results for each dataset is as follow:

| ALL-AML | Lung cancer | Prostate cancer | DLBCL |
|---|---|---|---|
| 58.88% | 74.02% | 56.99% | 98.0% |

These results are not as good as we were expecting. This method works perfectly for the DLBCL dataset however.

One reason for not getting perfect accuracy with this approach might be the nature of expression data.

*Using same threshold for all genes:* We tested the method using the third threshold selection technique, namely, selecting the threshold from the set {–1024, –512, …., –2, 0, 2, 4, 0……, 512, 1024}. Once a threshold is found for a dataset, it will be used for all genes within that set. This method provides very good accuracy in sample classification. Table 2 summarizes the results, each dataset was tested with the same threshold using the subsets of 1, 2,…,10 selected genes.

This approach provides results similar or superior to the previous reported results on this task [2]. Figure 1 illustrates our results compared with the previous published results taken from [2]. Figure 2 illustrates the classification accuracy for each threshold using only the best 10 genes selected by our method.

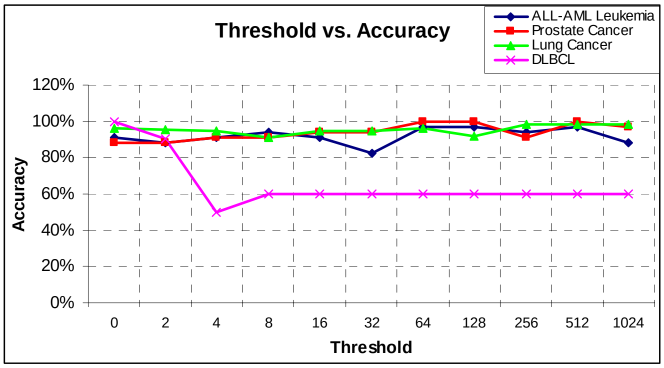


**Figure 2:** Illustration of the accuracy versus thresholds between 0 and 1024, the subset size in all datasets is 10 genes.

From the results of the three threshold techniques, we notice that the best accuracy for each dataset can be achieved with a different number of genes. We conducted another group of experiments to determine how many genes are needed for achieving the best classification accuracy in each dataset, and the results are in Table 3. The results in Table 3 show for each dataset how many genes are needed to achieve the best accuracy at four different thresholds. The smaller subsets of selected genes are more preferable than the larger subsets. The experiments on our method are so far conducted using at most 50 genes. We would like to evaluate our method with a little larger subset size. For that, we examined the accuracy for each dataset with 50 and 100 most useful genes, and the detailed results are in Table 4. As we have seen in Table 3, our method's best results are produced with 50 or less most informative genes, and the results of 100 genes in Table 4 confirm this.

## 4. CONCLUSION

We presented a new technique for gene selection using thresholding techniques and the notions of a,b,c,and d values. The method extracts a small gene subset that allows for sample classification with high accuracy. Since the gene expression profiles are usually produced from disease and normal tissue samples, the extracted genes are considered as related with that underlying disease. Furthermore, identifying a small group of genes that can classify the gene expression data with very high accuracy leads to the discovery that these selected genes are *associated* that disease studied in the gene expression dataset. The method proposed in this paper produces very promising and competitive results compared with the recently published results on this task. We have demonstrated the efficiency of the method with an extensive evaluation using four different datasets used mainly and extensively for this purpose. In the future work, we want to explore the ways to find, for each gene separately, the best threshold that yields the highest *v* value for that gene which will further improve the classification accuracy. Moreover, we plan to investigate how can we support our findings by exploiting the huge amounts of biomedical literature (*e.g. Medline*) using text-mining techniques.

| | AMLALL | | Prostate Cancer | | Lung Cancer | | DLBCL | |
|---|---|---|---|---|---|---|---|---|
| | No. of Genes | Accuracy | No. of Genes | Accuracy | No. of Genes | Accuracy | No. of Genes | Accuracy |
| **Threshold=0** | 50 genes | 97.06% | 8 genes | 91.18% | 1 gene | 97.32% | 3 genes | 100% |
| **Different Thresholds for Different Genes** | 50 genes | 88.24% | 3 genes | 73.53% | 9 genes | 99.33% | 2 genes | 100% |
| **Threshold=64** | 6 genes | 97.06% | 2 genes | 100% | 2 genes | 96.64% | 50 genes | 70% |
| **Threshold=512** | 4 genes | 97.06% | 2 genes | 100% | 4 genes | 99.33% | 50 genes | 70% |

**Table 3:** The best gene subset size and accuracy for each dataset in all experiments.

| Dataset | | Threshold=0 | Different Threshold for Different Genes | Threshold=64 | Threshold=512 |
|---|---|---|---|---|---|
| | | ***Accuracy*** | ***Accuracy*** | ***Accuracy*** | ***Accuracy*** |
| **AMLALL** | 100 genes | 64.71% | 73.53% | 85.29% | 67.65% |
| | 50 genes | 97.06% | 88.24% | 85.29% | 85.25% |
| **Prostate Cancer** | 100 genes | 88.24% | 73.53% | 100.00% | 97.06% |
| | 50 genes | 76.47% | 73.53% | 100.00% | 100% |
| **Lung Cancer** | 100 genes | 89.93% | 89.93% | 89.93% | 93.29% |
| | 50 genes | 89.93% | 93.29% | 89.93% | 95.97% |
| **DLBCL** | 100 genes | 100.00% | 80.00% | 70.00% | 70% |
| | 50 genes | 100.00% | 60.00% | 70.00% | 70% |

**Table 4:** The accuracies for all datasets with 50 and 100 genes.

## 5. REFERENCES

[1] T. K. Paul and H. Iba. Extraction of Informative Genes from Microarray Data. GECCO’05, June 25–29, 2005, Washington, DC, 2005.

[2] B. Liu, Q. Cui, T. Jiang and S. Ma. A combinational feature selection and ensemble neural network method for classification of gene expression data. BCM Bioinformatics 5:136, 2004,.

[3] G. J. Gordon, R. V. Jensen, L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker and R Bueno. Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma. CANCER RESEARCH 62, 2002.

[4] T. N. Do and F. Poulet. Incremental SVM and Visualization Tools for Bio-medical Data Mining. Proceedings of the European Workshop on Data Mining and Text Mining for Bioinformatics, 2003.

[5] L. Shen and E. C. Tan. Dimension Reduction-Based Penalized Logistic Regression for Cancer Classification Using Microarray Data. IEEE/ACM Transactions on computational biology and bioinformatics, 2 (2), 2005.

[6] T.R. Golub, DK. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, JP. Mesirov, H. Coller, M. L. Loh, J. R. Downing, MA. Caligiuri, CD. Loombeld, ES. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science Vol. 286, October 1999.

[7] A. A. Alizadeh, MB. Eisen, RE. Davis, C. Ma, IS. Lossos, A. Rosenwald, JC. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, GE. Marti, T. Moore, JH. Jr, L. Lu, DB. Lewis, R. Tibshirani' G. Sherlock, WC. Chan, TC. Greiner, DD. Weisenburger, JO. Armitage, R. Warnke, R. Levy, W. Wilson, MR. Grever, J. C. Byrd, D. Botstein, PO. Brown and LM. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature vol.403 Feb. 2000.

[8] Y. Lee and C. K. Lee. Classification of multiple cancer types by multi category support machines using gene expression data. Bioinformatics 19(9), Oxford University Press 2003.

[9] M. Kuramochi and G. Karypis. Gene Classification using Expression Profiles: A Feasibility Study. WSPC/Instruction File, March 18, 2005.

[10] D. V. Nguyen and D. M. Rocke. Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics 18(9), Pages 1216-1226, Oxford University Press 2002.

[11] D. Chaussabel and A. Sher. Mining microarray expression data by literature profiling. Genome Biology 3(10), 2002.

[12] S. Dudoit, J. Fridlyand and T. P. Speed. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Technical report # 576, June 2000.

[13] J. Newton. Analysis of Microarray Gene Expression Data Using Machine Learning Techniques .CMPUT 606: Computational Molecular Biology and Bioinformatics.

[14] H. Liu, J. Li and L. Wong. A Comparison Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns. Genome Informatics 13: 51-60, 2002.

[15] V. Vapnik. The nature of statistical learning theory. Springer, New York, USA, 1995.

[16] B. E. Boser, I. Guyon, V. Vapnik. A training algorithm for optimal margin classifiers. In COLT, pages 144-152, 1992.

[17] Kent Ridge biomedical datasets, univ http://sdmc.lit.org.sg/GEDatasets/Datasets.html