THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Efficient K-means Clustering with Greedy Algorithm for Minimum Gene Set Identification in Microarray Data

V. Devi
HOD and Assistant Professor, Department of Computer Applications
Guru Nanak College, Chennai, India
S. Natarajan
HOD and Assistant Professor, Department of Plant Biology and Biotechnology
Guru Nanak College, Chennai, India

Abstract:
Microarray data analysis is a widely used technique to unravel the gene expression and the ultimate aim is to find the differentially expressed genes in a particular condition. The initial step in a micro array data analysis is to identify the minimum subset of genes that may be differentially expressed, since the target number of genes in a micro array data is thousand of genes that are expressed together. Analyzing thousands of genes from a particular experiment is needless, as only few genes are involved in a particular disorder/dysfunction. Hence, the present research work aims at developing an efficient K-means clustering with greedy algorithm for identification of minimum subset of genes that may be differentially expressed in a micro-array experiment. It was found that the developed algorithm is more effective than the existing methods of gene subset selection.

***Keywords**: Gene Subset, Microarray, Clustering*

1. Introduction

The advent of DNA microarrays have resulted in There is an increasing acquisition of biological data in recent years due to advancements in molecular genetics technologies such as DNA Microarrays. [1-8]. The DNA Microarrays also called DNA Chips have given a global view of a cell for the first time. It is now possible to determine the molecular state of a cell due to simultaneous expression of thousands of genes using DNA microarrays. But, the challenge lies in the analysis of microarray data to interpret meaningful results. It is due to the fact that microarray data consists of small number of records with number of fields (i.e., number of genes) is in thousands which is in contrast to other data mining applications. This leads to a high likelihood of "false positives" in finding differentially expressed genes. The main types of data analysis needed to for biomedical applications include: Gene Selection – in data mining terms this is a process of attribute selection, which finds the genes most strongly related to a particular class ([9-15]), Classification – classifying diseases or predicting outcomes based on gene expression patterns, and perhaps even identifying the best treatment for given genetic signature ([21-31]) and Clustering – finding new biological classes or refining existing ones ([16-20]).

Since only a small number of genes among tens of thousands show strong correlation with the targeted disease, some works address the problem of defining which the appropriate number of genes to select is. However, finding the optimal number of genes, is a difficult goal. While an excessively conservative estimate of the number of relevant genes may cause an information loss, an excessively liberal estimate may increase the noise in the resulting dataset. This paper describes a method which automatically selects the minimum number of genes to reach the best classification accuracy on the training set. Moreover, the genes belonging to the minimum subset can be used as genetic markers for further biomedical investigations. Two different strategies to select the minimum set of genes which maximize the training samples coverage are exploited. The performance of our method has been tested on publicly available datasets and experimentally compared to other feature selection algorithms.

2. Pre-Processing

A microarray dataset (M) is represented as a gene expression matrix, with each gene representing a row and column corresponding to sample. Each element (m_{ij}) in the matrix is the expression level of all the genes under study and each e_{ij} in E is the expression level of gene 'i' for sample 'j' where i=1…..N; j=1…….M (N is the number of genes and M is the number of samples). In a microarray data, each sample is associated with a class label which can be either the clinical condition of the patient or the biological state of the tissue.

Each sample is also characterized by a class label, representing the clinical situation of the patient or the biological condition of the tissue. The domain of class labels is characterized by C different values and label ls of sample sn takes a single value in this domain.

$$E= \begin{bmatrix} e11\ e12 & & e1s \\ E21\ e22 & & e2s \\ & & \\ En1\ en2 & & ens \end{bmatrix}$$

According to this representation we define:

- Core expression interval. Definition of the range of expression values for a given gene in a given class. Two different approaches are exploited in the core expression interval definition.
- Gene mask. A new representation, named *gene mask*, captures the capability of a gene in distinguishing the sample classes (i.e., classification power) is introduced. Definition of the *gene mask* as representatives of gene classification power, where the classification power is the capability of a gene in discriminating the sample classes. The gene mask is generated by analyzing the gene core expression intervals. These definitions will be used in the following chapters to identify the genes which have a high discriminative power among classes in order to improve classification accuracy and to evaluate the similarity among groups of genes under different experimental conditions (i.e., sample classes).

The format of a microarray dataset conforms to the normal data format of machine learning and data mining, where a gene can be regarded as a feature or attribute and a sample as an instance or a data point. However, the main characteristics of this data type are the high number of genes (usually tens of thousands) and the low number of samples (less than one hundred). This peculiarity causes specific challenges in analyzing microarray data (e.g., complex data interactions, high level of noisy, lack of biological absolute knowledge) which have to be addressed by data mining methods. In recent years an abundance of microarray datasets become public available due to the increase of publication in bioinformatics domain. A large collection of public microarray data is stored by the ArrayExpress archive (http://www.ebi.ac.uk/microarray-as/ae/). The datasets, stored in MIAME and MINSEQE format, are all preprocessed, but also the raw data (for a subset of the collection) can be downloaded. One of the best features of this archive is the possibility to browse the entire collection or perform queries on experiment properties, submitter, species, etc. In the case of queries, the system retrieves summaries of experiments and complete data. Other datasets can be also downloaded from the author or tool websites (e.g., LibSVM software, GEMS software).

3. Methodology

The aim of identifying the minimum number of genes that is relevant for a particular target disease and is useful for correct classification of the samples in the data set. The method works in three phases:

- Class interval definition: The class expression interval for gene i and class k is expressed in the form:

$$I_{i,k} = [\min_{i,k}, \max_{i,k}] \qquad \text{-------------------} \ 1$$

where mini,k and maxi,k are the minimum and the maximum gene expression values for class k.

- Gene mask computation. For each gene we define a gene mask, which is an array of S bits, where S is the number of samples. It represents the capability of the gene to classify correctly each sample, i.e., its classification power. Consider an arbitrary gene i. Bit s of its mask is set to 1 if the corresponding expression value e $_{is}$ belongs to the core expression interval of a single class, otherwise it is set to 0. Formally, given two arbitrary classes c1, c2 ∈ C = {1, . . . , k}, bit s of gene mask i is computed as follows.

$$mask_{is} = \begin{cases} 1 \text{ if } (e_{is} \in Ii,c1) \mid e_{is} \in Ii,c2 \\ 0 \text{ otherwise} \end{cases} \qquad \text{-----------------} \ 2$$

A sample might not belong to any core expression interval (i.e., it is an outlier). In this case, the value of the corresponding bit is set to 0 according to 2. The gene mask shows which training samples the gene can unambiguously assign to the correct class. It is a string of 0s and 1s, generated by analyzing the overlaps among the class expression intervals (i.e., the range of expression values of samples belonging to the same class) for each gene.

- Minimum gene subset selection. The minimum number of genes needed to provide the best training set sample coverage is selected by analyzing the gene masks and exploiting the overlap scores. For this purpose two different searching algorithms (i.e., improved greedy approach and set covering) are exploited.

The minimum subset of genes for classifying the maximum set of samples in the training set (i.e., for providing the best sample coverage) and avoiding redundant information is defined by analyzing the gene masks.

Let G be a set of genes. We define a *global mask* as the logic OR between all the gene masks belonging to genes in G. The objective is the definition of the minimum set of genes G that holds enough discriminating power to classify the maximum number of samples in the training set. Thus, given the gene mask of each gene, we search for the global mask with the maximum number of ones. To this aim, we propose two different techniques: an improved greedy approach and a set covering approach. In our proposed improved greedy approach we have employed, K means clustering approach to cluster the data and greedy algorithm for classification.

4. K means Clustering with Greedy approach

The improved greedy approach takes the number of preferred outliers as input and selects points as outliers in a greedy approach. At first, the set of outliers (represented by OS) is specified to be empty and all points are represented as non-outlier. Subsequently, scans are necessary over the dataset to choose points as outliers. In every scan, for each point labeled as non-outlier, it is temporally removed from the dataset as outlier and the entropy object is re-evaluated. A point that accomplishes maximal entropy impact, i.e., the maximal reduction in entropy experienced by removing this point, is taken as outlier in current scan and accumulated in OS. The algorithm ends when the size of OS reaches the total number of genes.

This approach identifies at each step the gene with the best complementary gene mask with respect to the current global mask. Thus, it adds at each step the information for classifying most currently uncovered samples. The pseudo-code of our improved Greedy approach is reported in Algorithm 1. It takes as input the set of gene masks (M), the set of scores (OS) and produces as output the minimum subset of genes (G). The scores associated to each gene can be computed by different feature selection techniques. The variance of expression values is exploited as score. The first step is initializing G at Ø (line 2), the candidate set (C) at Ø (line 3), and clustering of gene data (lines 5 to 9). Initially, all the global mask is set to zeros (line 10). Then the following steps are iteratively performed.

Algorithm : Minimum gene subset –
K means clustering with Greedy approach

```
Input: set M of all the mask $_i$, set OS of score os $_i$ for each gene i
Output: set G of genes
1: /*Initialization*/
2: G = Ø
3: C = Ø /*candidate gene set at each iteration*/
4: /*Clustering*/
5: For k=1: length(D)                /* D is the input */
6: If k==1
7: O=1                         /* identified clusters*/
8: b
9: k=0
10: global mask = all_zeros() /*vector with only 0s*/
11: /*Control if the global mask contains only 1s*/
12: while not global_mask_all_ones() do
13: /*Determine the candidate set of genes with most ones*/
14: C = max_ones_genes()
15: if C != Ø then
16: /*Select the candidate with the best score (e.g., the minimum)*/
17: c = C[1]
18: for all j in C[2 :] do
19: if $OS_j$ is better $OS_c$ then
20: c = j
21: end if
22: end for
23: /*Update sets and global mask*/
24: G = G + c
25: global mask = global mask OR maskc
26: M=M− maskc
27: /*Update the masks belongs to M*/
28: for all maski in M do
29: $mask_i$ = $mask_i$ AND global mask
30: end for
31: else
32: break
33: end if
34: end while
35: return G
```

- The gene mask with the highest number of bits set to 1 is chosen (line 14). If more than one gene mask exists, the one associated to the gene with the best score is selected (lines 15-22). The best score depends on the range values produced by the technique exploited. For example, if the variance is used as score the gene with the highest variance is selected.
- The selected gene is added to set G (line 24) and the global mask is updated by performing the logical OR between the gene mask and the global mask (line 25).
- The gene masks of the remaining genes (gene mask set M, line 26) are updated by performing the logical AND with the negated global mask (lines 27-30). In this way, only the ones corresponding to the classification of still uncovered samples are considered.
- If the global mask has no zeros (line 12) or the remaining genes have no ones (line 15), the procedure ends.

5. Set Covering Approach

The set covering approach considers the set of gene masks as a matrix of N × S bits and performs the following three steps.

- *Sample reduction*. Each sample (i.e., column) that contains all 0 or 1 over the N gene masks is removed, because it is uninformative for the searching procedure.
- *Gene reduction.* Each gene (i.e., row) whose gene mask is a subsequence of another gene mask is removed from the matrix. If two or more genes are characterized by the same gene mask, only the gene with the best score is kept in the matrix. At the end of these two steps a reduced matrix is obtained.
- *Reduced matrix evaluation*. The reduced matrix is evaluated by an optimization procedure that searches the minimum set of rows necessary to cover the binary matrix. Since it is a min-max problem, it can be converted to the following linear programming problem.

$$\text{Min} \quad \sum_{i=1}^{i=N} g_i$$

$$\sum_{i=1}^{i=N} mask_{ij}.g_i >= 1, \; j=1,..,S$$

$$gi \in \{0, 1\}$$

The branch and bound implementation provided by the Symphony library [116] has been exploited to find the optimum solution. At the end of this phase, the minimum set of genes required to provide the best sample coverage of the training set is defined. The genes in the minimum subset are ordered by decreasing number of 1s in the gene mask.

6. Experimental Results

We validated our method by comparison with other feature selection techniques on public gene expression datasets. Classification accuracy is used as the performance metric for evaluation, while biological relevance of the selected genes is also discussed.

- **Data Set:** We evaluated the performance of our algorithm on four microarray datasets, publicly available [32]. Two of them are multi-class (Brain1 and Brain2), and the other two are binary (SRBCT and DLBCL). Table 1 summarizes their characteristics.

Dataset	Samples	Features	Classes
Brain1	90	5920	5
Brain2	60	10364	4
SRBCT	83	2308	2
DLBCL	77	5469	2

Table 1: Dataset characteristics

7. Experimental Setting

We compared the performance of our method with the following supervised feature selection methods implemented in RankGene software: Information Gain (IG), Twoing Rule (TR), Sum Minority (SM), Max Minority (MM), Gini Index (GI), Sum of Variance (SV). For each of these methods we performed experiments with a number of selected features in the range from 1 to 12 to allow the comparison with the number of features selected by our method. We exploited the libSVM classifier [33] with a 4-fold cross validation. Samples were randomly partitioned in a stratified manner into four folds of equal size. Three folds become the training set and the fourth the test set. Classification is repeated for four times, each time with a different fold as test set. To avoid the selection bias, feature selection algorithms were applied only to the training set, and accuracy was computed by applying the classifier to the test set. We repeated the cross-validation 50 times, changing the seed for the split generation.

8. Classification Accuracy

As shown in Table 2, the gene reduction step significantly reduces the number of considered features. In the second column the average value of remaining features over 50 repetitions of the 4-fold cross validation is reported. The reduction rate (i.e. the number of discarded features over the total number of features) in the third column highlights that there are between 60% and 90% genes with a gene mask which is a subsequence of another gene mask. This affects the number of genes selected by our algorithm. For example, Brain2 and DLBCL, which have the highest reduction rates, also have the minimum number of selected genes, as shown in the fourth column of the table. This average value is always slightly less than the average number of features selected by the greedy approach reported in the fifth column of the table.

The average accuracy over 50 repetitions of the 4-fold cross validation is reported in Figures 4.1, 4.2, 4.3 and 4.4, where the performance of the mask covering algorithm and the six RankGene methods are compared. Also the accuracy of the greedy approach is reported. For the mask covering and the greedy algorithm a single point is shown on the graph, because these methods automatically select the best number of features for the dataset. Instead, for the other methods, the behavior varying the gene number may be analyzed, as they require the gene number as input parameter.

For the DLBCL dataset (Figure 1) our method needs only 3 or 4 genes, depending on the number of samples in the split (the diagram reports the average gene number value over all repetitions), to reach an accuracy higher than 88%, while other methods do not reach such accuracy even when using 10 or more genes. Also the greedy approach performs slightly worse than our method, even if the number of selected genes is higher. For the Brain2 dataset (Figure 2) our method reaches a high accuracy (about 63%) with less than 5 genes, while the other methods reach a lower accuracy also with a higher number of genes.

For the SRBCT dataset (Figure 3) our method selects the best small subset of genes. However classification accuracy may be further improved by increasing the cardinality of the gene set. Finally, in the Brain1 dataset (Figure 4) our method is less effective in detecting a very small set of high quality features and shows a behavior closer to other feature selection techniques. These differences in the accuracy levels depend on the dataset characteristics.

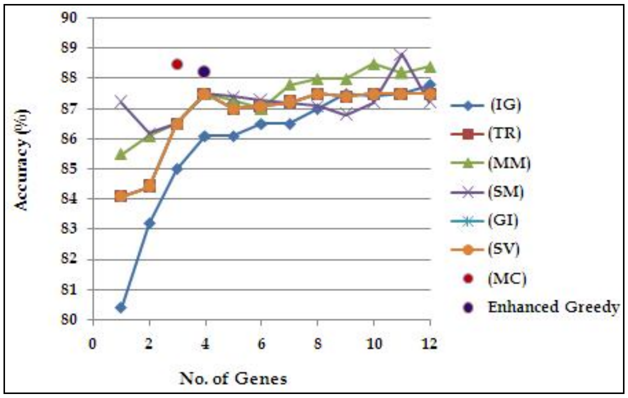
The genes selected by each method are mostly different from the ones selected by the other methods. The percentage of common genes is about 10-20%. It means that there are a lot of genes which classify same samples and each method select different genes to reach a similar accuracy. We performed the Student's t-test to assess the statistical significance of the results. We compared our method with each of the RankGene methods by setting as gene cardinality the integer number nearest to the mean value of genes selected by our method. We obtained a p-value less than 0.01 in 3 over 4 datasets (Brain2, SRBCT, DLBCL). In the case of Brain1 the p-value value is less than 0.05 for some methods.

Dataset	Remaining features	Reduction rate	Mask Covering	Improved Greedy
Brain1	1874	68%	6.76	7.80
Brain2	847	92%	4.62	5.05
SRBCT	660	71%	5.28	5.75
DLBCL	1245	77%	3.50	3.79

Table 2: Reduction rate and average number of selected features

9. Biological Discussion

We investigated the biological meaning of our selected genes. For the DLBCL dataset, the genes selected by the mask covering algorithm include the Tcell chemoattractant SLC and the DNA replication licensing factor CDC47 homolog, which are known to be related to lymphoma. Furthermore, the genes selected by the greedy approach include the DNA replication licensing factor CDC47 homolog, the Cancellous bone osteoblast mRNA for GS3955 and the Chloride channel (putative) 2163bp, which are listed as relevant for lymphoma.



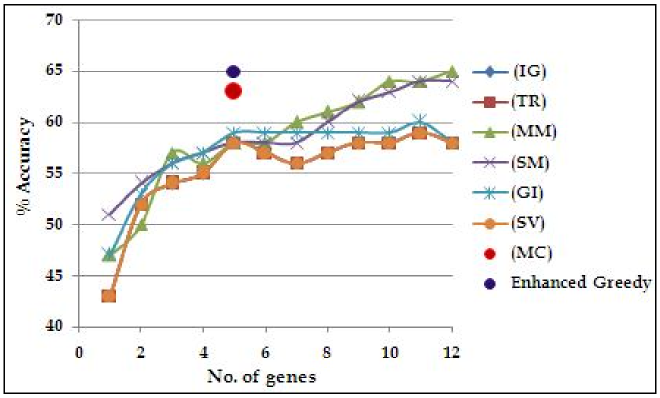
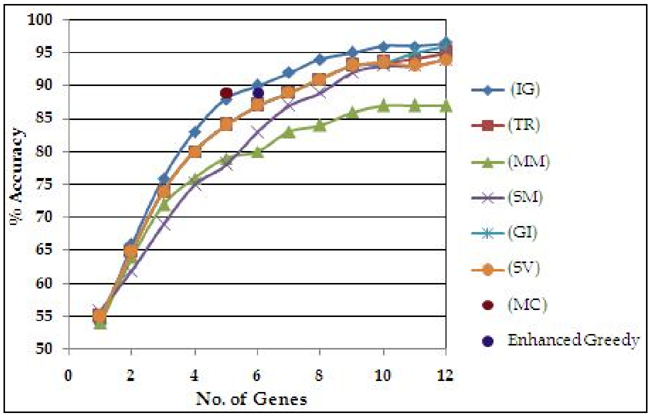


Figure 1: Mean Classification accuracy of six RankGene methods, Mask Covering and Enhanced Greedy on DLBCL dataset
Figure 2: Mean Classification accuracy of six RankGene methods, Mask Covering and Enhanced Greedy on Brain2 dataset



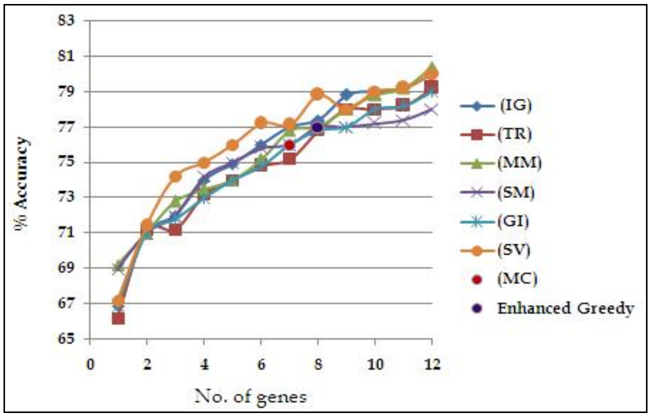


Figure 3: Mean Classification accuracy of six RankGene methods, Mask Covering and Enhanced Greedy on SRBCT dataset
Figure 4: Mean Classification accuracy of six RankGene methods, Mask Covering and Enhanced Greedy on Brain1 dataset

10. Conclusion

The proposed work has clearly established a better performance than any of the existing methods. The method selects the minimum number of genes required to a high accuracy level by using a new representation of the genes called mask covering to differentiate among classes. The usage of set covering approach has identified a small number of genes that have biological significance also. Thus the initial step of microarray data analysis, the selection of minimum number of differentially expressed genes has become more productive.

11. References

1. Chipping Forecast 1999, 2002, The Chipping Forecast. Special Supplement. Nature Genet. 21, Jan. 1999.
2. The Chipping Forecast II. Special Supplement. Nature Genet. 32, Dec. 2002
3. Schena, M. et al Quantitative monitoring of gene expression patterns with a cDNA microarray. Science 270:467-470 (1995).
4. DeRisi, J.L. et al. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278: 680-686 (1997).
5. Chu, S. et al. The transcriptional program of germ cell development in budding yeast. Science 282:699-705 (1998).
6. Iyer, V.R. et al. The transcriptional program in the response of human fibroblasts to serum. Science 283: 83-87, (1999)
7. DeRisi J, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996 Dec;14(4):457-60.
8. Hegde P. et al. A concise guide to cDNA microarray analysis. Biotechniques. 2000 Sep; 29(3):548-50, 552-4, 556.
9. Marchal K et al Comparison of different methodologies to identify differentially expressed genes in two-sample cDNA microarrays. JOURNAL OF BIOLOGICAL SYSTEMS 10 (4): 409-430 DEC (2002).
10. Baldi P and AD Long. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics, 17: 509-519, (2001).
11. Li C and WH Wong. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biology, (2001)/2/8/research/0032.
12. Tusher VG et al. Significance analysis of microarrays applied to the ionizing radiation response. PNAS, 98:5116-5121, (2001).
13. Dudoit S et al. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica, 12:111-139, (2002).
14. Ideker, T. et al. Testing for differentially-expressed genes by maximum likelihood analysis of microarray data. Journal of Computational Biology, 7, 805-817 (2000).
15. Storey J. D. and R. Tibshirani. Statistical significance for genome wide studies. PNAS, August 5, 2003; 100(16): 9440 – 9445 (2003).
16. Eisen M. et al. Cluster analysis and display of genome-wide expression patterns. PNAS, 95:14863-14868 (1998).
17. Tamayo P. et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. PNAS, 96:2907-2912, (1999).
18. Hastie T. et al. Supervised harvesting of expression trees. Genome Biology, 2(1) :research0003.1-0003.12, (2001).
19. Li H and F. Hong. Cluster-Rasch models for microarray gene expression data. Genome Biology, 2(8)}:research0031.1-0031.13, (2001).
20. Lin W. and C. Le Model-based cluster analysis of microarray gene expression data. Genome Biology, 3(2): research0009.1-0009.8, (2002).
21. Golub T. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286:531-537, 1999.
22. Alizadeh L. et al. Identification of clinically distinct types of diffuse large B-cell lymphoma based on gene expression patterns. Nature 403: 503-511 (2000).
23. Bittner M. et al. Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling. Nature 406: 536-540 (2000)
24. Ramaswamy S. et al. Multi-Class Cancer Diagnosis Using Tumor Gene Expression Signatures, PNAS 98: 15149-15154.
25. Tibshirani R, et al. "Diagnosis of multiple cancer types by shrunken centroids of gene expression" PNAS 2002 99:6567-6572 (May 14).
26. Ramaswamy S. et al. Evidence for a Molecular Signature of Metastasis in Primary Solid Tumors. Nature Genetics, vol. 33, January 2003, pp. 49-54.
27. Khan J. et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, Volume 7, Number 6, June 2001.
28. Hedenfalk I. et al. Gene Expression Profiles in Hereditary Breast Cancer. NEJM, 244:539-548. (2001).
29. Chang HY et al. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. PLoS Biol. 2004 Feb; 2(2): 1.
30. Nutt CL. Et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res. 2003 Apr 1;63(7):1602-7.
31. Lapointe J. et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. PNAS 2004 Jan 20; 101(3): 811.
32. A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631, 2005.

33. C.C. Chang and C.J. Lin. Training v-support vector classifiers: theory and algorithms, 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm