Weighted Top Score Pair Method for Gene Selection and Classification

Huaien Luo[1], Yuliansa Sudibyo[2,⋆], Lance D. Miller[1],
and R. Krishna Murthy Karuturi[1,⋆⋆]

[1] Genome Institute of Singapore, Singapore
[2] Nanyang Technological University, Singapore
{luoh2,gisv45,millerl,karuturikm}@gis.a-star.edu.sg

Abstract. Gene selection and expression profiles classification are important for diagnosing the disease using microarray technology and revealing the underlying biological processes. This paper proposes a weighted top scoring pair (WTSP) method which is a generalization of the current top scoring pair (TSP) method. By considering the proportions of samples from different classes, the WTSP method aims to minimize the error or misclassification rate. Results from several experimental microarray data have shown the improved performance of classification using the WTSP method.

Keywords: Microarray, Gene selection, Classification, Weighted Top Score Pairs, Cross-validation.

1 Introduction

By measuring the expression levels of thousands of genes, microarray techniques have been used to diagnose and explore the biologically relevant genes related to a disease. The obtained microarray data normally contains several thousands of genes and tens to hundreds of samples. The analysis of this data is challenged by the "small N, large P" problem, that is, the number of genes (P) is greatly larger than the number of samples (N). In order to deal with this high dimensional data and make the analysis feasible, dimension reduction (or gene selection) methods are used to choose the most informative genes by comparing the expression levels between the cancer tissue and normal ones, or between different tumor types. The purpose of the gene selection is to discard those genes which are least interesting to the classification and select the relevant genes which could provide the best ability to distinguish the samples from different classes and hence reveal the biomarkers or molecular signature for the disease. This purpose can be achieved by ranking the genes according to some relevance measurement and select those genes with the highest relevance scores. The commonly used genes selection methods can be categorized into three categories: i) choosing

[⋆] This author is the co-first author.
[⋆⋆] Corresponding author.

M. Chetty, A. Ngom, and S. Ahmad (Eds.): PRIB 2008, LNBI 5265, pp. 323–333, 2008.

single differentially expressed genes; ii) choosing gene pairs which co-regulate; and iii) choosing a set of genes or gene network. To discover the differentially expressed genes, t-statistic could be calculated for each gene and the genes with significantly different expression levels are chosen [1]. This single gene selection method considers the genes independently and may miss the functional relationships among genes due to the interaction/co-regulation of the genes. Some methods are proposed to investigate the information provided by the gene pairs. In [2], two-sample t-statistics are calculated for each gene pairs projected to the diagonal linear discriminant (DLD) axis in order to find the pairs with the highest score that together could discriminate the samples from different classes. In [3], a correlation-based method is developed to discover the gene pairs whose functional relationship changes across different conditions. This method is based on the assumption that gene pairs with largest differential correlation are more likely to be involved in the mechanisms of the disease. In [4], a feature construction method is proposed to find the synergic gene pairs which could enhance the accuracy of the classification. In this method, the mutual information contained in the interaction of the gene pairs is explored for the gene selection and these gene pairs are assumed to have biological significance for the underlying cellular processes. In addition to investigate the information contained in pairs of genes, the microarray data analysis can also be carried out with a list of genes (or gene networks). The information buried in this gene network could reveal the biological function or pathways of these genes related to the disease. In [5], the gene expression data is analyzed by integrating a priori the knowledge of the gene network to achieve a better classification. The hypothesis underlying this approach is that the genes close to the network are more likely to be co-expressed. In [6], a friendly neighbors (FNs) method for time-course microarray data analysis is proposed to find the genes whose induction-repression pattern are shared with other genes more often and these genes are considered to be the most informative for a certain cellular function. Based on this method, a differential friendly neighbors (DiffFNs) method is proposed to choose the genes in which the gain or loss of the relationships with other genes are most significant [7]. These genes could provide the biomarkers to distinguish the tumor from the healthy ones and signify the underlying pathways.

Besides the above methods, the common dimension reduction methods are also used to represent the information of the large number of genes with a set of gene components which could capture as much information of the original gene expression data as possible. These methods include: Q-mode Principle Component Analysis (PCA) which retain most of the variation [8]; Partial Least Squares (PLS) which constructs the components that maximize the covariance between classes and genes [9]; Sliced Inverse Regression (SIR) which regresses the gene expression data on the classes [9].

Having selected the most informative genes, the samples from different classes could be successfully identified. Many algorithms have been proposed to achieve this goal, such as Support Vector Machine (SVM) [10], nearest and k-nearest neighbors (kNN) [11][12], linear discriminant analysis (LDA) [11], Decision Trees

(DT) [13], naive Bayes (NB) [11], Prediction Analysis of Microarrays (PAM) [14] and so on.

Among these gene selection and classification methods, one of the simple and effective methods is Top Scoring Pair (TSP) based methods [15][16][17]. This method integrates the gene selection and classification based on a simple rule. It aims to find pairs of genes such that the expression level of gene A is greater than that of gene B in class 1, but smaller in class 2; and this rule is also used for the classification. Being a rank-based method, the TSP is invariant to the preprocessing steps such as normalization since it does not change the rank of a specified gene. Compared to the traditional methods which use more genes and a complex decision procedure, the TSP method is shown to have the ability to achieve comparably high accuracy of classification by using very few genes [18].

In this paper, a weighted TSP (WTSP) method is proposed as a generalization of the classical TSP method. Different from the TSP, the proposed WTSP method adjusts the scores of gene pairs by incorporating the information of the proportion of the samples belonging to different classes and/or the cost of misclassification. This weighted TSP method aims to minimize the cost of misclassifications and hence could achieve better performance compared to the classical TSP. This paper is organized as follows. In Section 2, the method of weighted TSP will be developed. Some implementation issues will also be given in this part. Section 3 presents the results of the proposed method as well as its comparison with the TSP classifier. This is followed by Section 4 where some discussion about the proposed method will be given.

2 Method

The gene expression data can be represented as a matrix $\mathbf{X}$ with dimension $P \times N$, where P is the number of genes and N is the number of samples (or gene expression profiles). Each column in $\mathbf{X}$ is an expression profile of P genes from a sample either in class 1 ($Y = 1$) or in class 2 ($Y = 2$). Normally, the number of genes is greatly larger than the number of samples ($P \gg N$) and this causes the problem of curse of dimensionality.

The TSP method aims to find the gene pairs whose relative relationship of expression levels change from one class to the other. That is, the marker gene pairs should be the ones that the expression level of gene A is greater than that of gene B in class 1, but smaller in class 2. Suppose there are N_1 samples from class 1 and N_2 samples from class 2 ($N_1 + N_2 = N$), and for a gene pair (i, j), there are respectively a_{ij} and b_{ij} samples from class 1 and class 2 with the expression level of gene i less than that of gene j (i.e., $X_i < X_j$). The TSP scheme order the gene pairs according to their scores defined as:

$$\begin{aligned}\varDelta_{ij} &= |P(X_i < X_j|Y = 1) - P(X_i < X_j|Y = 2)| \\ &= |p_{ij}(C1) - p_{ij}(C2)| \\ &\approx |\frac{a_{ij}}{N_1} - \frac{b_{ij}}{N_2}| \end{aligned} \tag{1}$$

By choosing the gene pairs which achieve the top scores in the training data, a new gene expression profile $\mathbf{x}'$ could be classified according to the relation of the expression level X'_i of gene i and X'_j of gene j (or the rank of these two genes) according to the following rule:
If $p_{ij}(C1) > p_{ij}(C2)$,

$$Y' = \begin{cases} 1, & \text{if } X'_i < X'_j \\ 2, & \text{o.w.} \end{cases}, \tag{2}$$

else if $p_{ij}(C1) \leq p_{ij}(C2)$,

$$Y' = \begin{cases} 2, & \text{if } X'_i < X'_j \\ 1, & \text{o.w.} \end{cases} \tag{3}$$

2.1 Weighted TSP Method

The proposed weighted TSP method is based on the classical TSP with the incorporation of the probabilities of the samples belonging to each class and the cost of misclassification. It aims to minimize the cost of misclassification, that is, to minimize the following equation:

$$\text{Cost} = P(\text{error}|Y = 1)P_1\lambda_1 + P(\text{error}|Y = 2)P_2\lambda_2, \tag{4}$$

where, $P_1 = P(Y = 1)$ and $P_2 = P(Y = 2)$ are respectively the probability of the samples coming from class 1 and class 2; λ_1 and λ_2 represent the cost it may induce if a sample is misclassified.

If we specify the classification rule as: if $X_i < X_j$, the sample is classified to class 1 ($Y = 1$); else if $X_i > X_j$, it is classified to class 2 ($Y = 2$); and if $X_i = X_j$, the sample is assigned to the class with higher probability. Let a_{ij} be the number of samples correctly assigned to class 1 under this classification rule (i.e., either $X_i < X_j$ or $X_i = X_j$ with $P_1 > P_2$) and b_{ij} be the number of samples incorrectly assigned to class 2, Eq. (4) could be reduced to:

$$\text{Cost} = \frac{N_1 - a_{ij}}{N_1}P_1\lambda_1 + \frac{b_{ij}}{N_2}P_2\lambda_2 \tag{5}$$

$$= P_1\lambda_1 - (\frac{a_{ij}}{N_1}P_1\lambda_1 - \frac{b_{ij}}{N_2}P_2\lambda_2) \tag{6}$$

It can be easily observed from the above equation that the minimization of the cost of misclassification is actually equivalent to the maximization of the quantity $\frac{a_{ij}}{N_1}P_1\lambda_1 - \frac{b_{ij}}{N_2}P_2\lambda_2$. Therefore, for each gene pair, a weighted score is calculated according to:

$$\Delta'_{ij} = \frac{a_{ij}}{N_1}P_1\lambda_1 - \frac{b_{ij}}{N_2}P_2\lambda_2. \tag{7}$$

Compared to the original score, the weighted score Δ'_{ij} is a generalization of the original score Δ_{ij} by considering the proportion of the samples in each class as well as the cost of misclassification. Here, we consider two special cases of this weighted score.

1. If $P_1\lambda_1 = P_2\lambda_2$, Δ'_{ij} is reduced to a scaled version of the score Δ_{ij} calculated in the classical TSP as shown in Eq. (1). It can also be seen that the original score does not consider the proportions of samples from each class and hence the maximization of the original score is equivalent to minimizing the sum of misclassification probabilities over two classes instead of the probability of total misclassification.
2. If $\lambda_1 = \lambda_2$, minimization of the cost of misclassification in Eq. (4) is actually the minimization of the probability of total misclassification (or the error rate).

By ordering the scores of each pair, the gene pair with the largest score is chosen as the marker gene pair to classify the samples. And for a new expression profile $\mathbf{x}'$, the classification rule now is:

$$Y' = \begin{cases} 1, & \text{if } X'_i < X'_j;\ \text{or } X'_i = X'_j \text{ and } P1 > P2 \\ 2, & \text{o.w.} \end{cases} . \tag{8}$$

It is to be noted that in the proposed WTSP method, the absolute sign is discarded compared to the original method and the classification rule is also accordingly simplified. This is because that in the weighted score Δ'_{ij}, the order of the genes in the pair is considered. That is, the scores of both the pair (i, j) and (j, i) are calculated and only the one which can achieve higher score is kept for further analysis. While in the original TSP, the order of the genes in the pair is not considered and hence the absolute sign is used and the classification rule depends on the relative value of $p_{ij}(C1)$ and $p_{ij}(C2)$.

In practice, several gene pairs may achieve the same top score. The original TSP method uses two schemes to deal with this situation: i) use all the top score gene pairs and a majority voting scheme to classify the test samples [17]; ii) find the rank of the genes in the pair and choose the pair whose rank difference of the two genes is largest as the marker gene pair for classification [15]. In the WTSP method, a different scheme is used. We treat the gene pairs whose scores are close to the top score as having the same power to classify the samples. This is because that the relative relationship of X_i and X_j may reverse due to noise and this may cause that the measured a_{ij} and b_{ij} are slightly different from the real ones (especially when the X_i and X_j are close to each other). Therefore, it is desirable to treat these gene pairs as potential pairs to be chosen for classifying the test samples. Among these gene pairs, the marker gene pair should have the property that their expression levels are most negatively correlated. And this marker gene pair is the one used in the proposed WTSP (w/ corr.) for classifying the test samples.

2.2 Cross-Validation

In this paper, leave-one-out cross-validation (LOOCV) is used to estimate the error or misclassification rate. For each sample in the available training data with known class, we select the gene pair and build the classifier from the remaining samples. The sample which is left out is treated as the test sample and the classification is made according to the classifier established from the remaining training samples. The classification accuracy is then calculated as the correct classification divided by the number of samples.

Due to large number of genes, an efficient algorithm to perform the cross-validation is desired. To achieve this, an accelerated cross-validation scheme is utilized based on the idea that the gene pairs which possess very low scores can be ignored since they never have the chance to be chosen as the top scoring pair (or top two scoring pairs in the proposed WTSP (w/ corr.) method) no matter which sample is left out in the process of cross-validation. This can be realized by calculating the lower bound and upper bound of the weighted scores based on all samples for each gene pair. The following steps describe this procedure.

1. For each gene pair (i, j), first calculate the weighted scores Δ'_{ij} according to Eq. (7) by using all the samples, note down respectively the a_{ij} and b_{ij}.
2. Calculate the lower and upper bound of the weighted score of gene pair (i, j) when one sample is left out. This can be done by calculating the following four terms:

$$\Delta^1_{ij} = \frac{a_{ij}}{N_1 - 1} P_1 \lambda_1 - \frac{b_{ij}}{N_2} P_2 \lambda_2, \text{ if the sample is from class 1 and } X_i > X_j$$
$$\Delta^2_{ij} = \frac{a_{ij} - 1}{N_1 - 1} P_1 \lambda_1 - \frac{b_{ij}}{N_2} P_2 \lambda_2, \text{ if the sample is from class 1 and } X_i < X_j$$
$$\Delta^3_{ij} = \frac{a_{ij}}{N_1} P_1 \lambda_1 - \frac{b_{ij}}{N_2 - 1} P_2 \lambda_2, \text{ if the sample is from class 2 and } X_i > X_j$$
$$\Delta^4_{ij} = \frac{a_{ij}}{N_1} P_1 \lambda_1 - \frac{b_{ij} - 1}{N_2 - 1} P_2 \lambda_2, \text{ if the sample is from class 2 and } X_i < X_j$$

   It can be easily observed that $\Delta^2_{ij} < \Delta^1_{ij}$ and $\Delta^3_{ij} < \Delta^4_{ij}$.
   So the lower bound is then:

$$\Delta^L_{ij} = \min(\Delta^2_{ij}, \Delta^3_{ij}), \tag{9}$$

   and the upper bound is:

$$\Delta^U_{ij} = \max(\Delta^1_{ij}, \Delta^4_{ij}). \tag{10}$$

3. Find the lower bound of the top score pair based on all the samples, and discard those gene pairs whose upper bound is less than the lower bound of the top score pair since for these gene pairs, their weighted score cannot become the top one no matter which sample is left out.

By using this scheme, a list of gene pairs $\mathcal{L}$ is obtained. In each LOOCV loop, only the gene pairs in the list $\mathcal{L}$ are investigated and the weighted scores are updated as well. This procedure greatly reduces the number of gene pairs that we need to investigate and hence largely increases the time and space efficiency of the cross-validation.

3 Evaluation and Results

The proposed weighted TSP method was then tested on the data available from the public database as well as from our own side. These data sets are respectively Leukemia [19], Colon [20], Lung [21], DLBCL [22], GCM [23], CNS [24], Prostate [25], p53 [26]. Table 1 gives a summary of these data sets, such as the number of genes measured (P), total number of samples (N) and the number of samples in each class.

Table 2 shows the comparison of the performance of the proposed weighted TSP and the original TSP. The classification accuracy is estimated using the LOOCV. For the WTSP method, we choose $\lambda_1 = \lambda_2$ to calculate the weighted scores. In this table, "WTSP (w/o corr.)" means that all the gene pairs with the same top weighted scores are used to classify the samples and the classification result is based on the majority voting strategy. "WTSP (w/ corr.)" means weighted TSP with the consideration of the cross-correlation of the expression levels of the gene pair. The gene pairs with the weighted score at least second to the top ones are chosen as the potential gene pairs for classification and only

Table 1. Description of the Data Sets

Data sets	# genes (P)	# total samples (N)	# samples by class (N_1/N_2)
Leukemia	7129	72	47 ALL / 25 AML
Colon	2000	62	40 Tumor / 22 Normal
Lung	12533	181	150 ADCA / 31 MPM
DLBCL	7129	77	58 DLBCL / 19 FL
GCM	16063	280	190 Tumor / 90 Normal
CNS	7129	34	25 Classic / 9 Desmoplastic
Prostate	12625	88	50 Normal / 38 Tumor
p53	44928	257	59 p53+ / 198 p53-

Table 2. Classification Accuracy for 8 Data Sets

Data Sets	WTSP (w/o corr.)	WTSP (w/ corr.)	TSP (w/o rank)	TSP (w/ rank)
Leukemia	95.83%	**97.22%**	93.80%	94.44%
Colon	91.13%	90.32%	91.13%	**91.94%**
Lung	**99.17%**	95.58%	**99.17%**	98.30%
DLBCL	97.40%	94.80%	**98.05%**	97.40%
GCM	77.5%	**84.64%**	75.40%	75.40%
CNS	**83.82%**	79.41%	**83.82%**	79.41%
Prostate	65.34%	**75.00%**	55.68%	54.55%
p53	**79.76%**	79.00%	76.65%	76.65%
Average	86.24%	**87.00**%	84.21%	83.51%
Std	11.76 %	8.63%	14.66%	14.98%
Min	65.34%	75.00%	55.68%	54.55%
Max	99.17%	97.22%	99.17%	98.30%

the one which is most negatively correlated is chosen as the marker gene pair to classify the samples. Similarly, “TSP (w/o rank)” and “TSP (w/ rank)” represent the original TSP method respectively either using all the gene pairs having the same original top scores with majority voting strategy, or choosing the one whose average rank difference is largest [15].

From this table, it is clearly seen that on average, the WTSP-based methods work better than the original TSP-based methods. For 4 out of 8 cases, both WTSP (w/o corr.) and WTSP (w/ corr.) outperform either the TSP with or without rank (respectively Leukemia, GCM, Prostate and p53). For the data sets of Lung and CNS, the WTSP (w/o corr.) performs as well as the best of the TSP-based methods. Only in the DLBCL and Colon case, the WTSP method works slightly worse than the TSP method, but the difference is not significant (respectively, 0.65% and 0.81% difference). An obvious observation is that when the classification accuracies of TSP-based methods are high, the performance of the WTSP-based methods are comparable to the TSP-based methods. However, when the classification accuracies of TSP-based methods are low (such as in

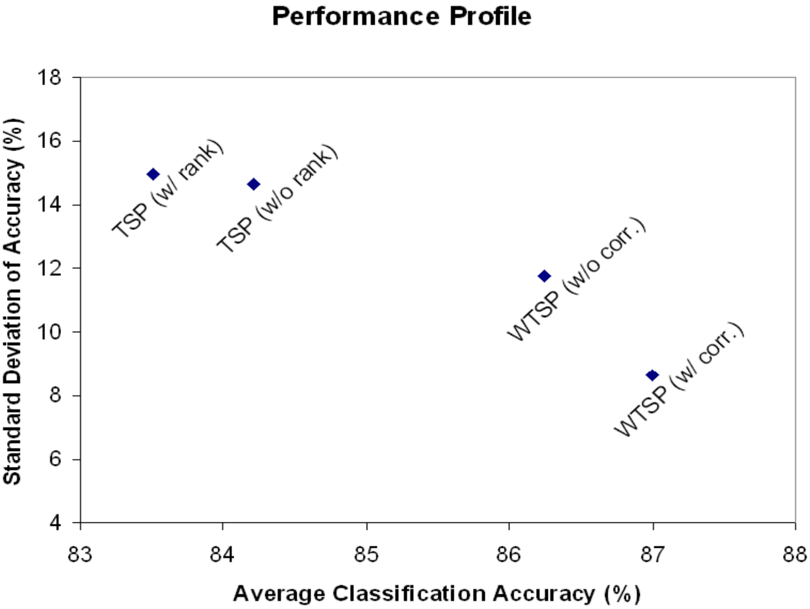

Fig. 1. Performance Comparison of WTSP and TSP methods. Each method is represented as a point in this figure with the coordinates composed of average classification accuracy and its standard deviation. The performance of WTSP (w/ corr.) is the best among these methods with highest average accuracy (87.00%) and smallest standard deviation (8.63%). The WTSP (w/o corr.) takes the second place with average accuracy (86.24%) and standard deviation (11.76%). The TSP-based methods perform worse compared to WTSP-based methods with lower average accuracies (84.21% for TSP without rank and 83.51% for TSP with rank) and relatively larger standard deviation (respectively, 14.66% and 14.98%).

GCM, Prostate and p53 data), the proposed WTSP-based methods significantly improve the performance. The increase of the accuracy are respectively 9.24% for GCM data, 19.32% for Prostate data and 3.11% for p53 data.

Figure 1 shows the performance comparison between the TSP-based and WTSP-based methods. The average accuracies of classification and their standard deviations for each method are calculated. The method which could achieve higher classification accuracy with lower standard deviation is desired (Ideally, a method with 100% accuracy and 0 standard deviation is the best). If the standard deviation of the accuracy is plotted against its average classification accuracy for each method as shown in this figure, it can be seen that when a method represented by a point in this figure is closer to the bottom-right corner, its performance will be better.

From this figure, it is clearly seen that the WTSP (w/ corr.) works best with the highest average accuracy of 87.00% and smallest standard deviation of 8.63%, followed by WTSP (w/o corr.), TSP (w/o rank) and TSP (w/ rank). The improvement of WTSP method comes from the fact that it minimizes the probability of total misclassification and hence it may choose different gene pairs than the ones chosen by the original TSP method for classification.

4 Discussion and Conclusion

Based on the classical TSP method, we proposed a weighted TSP (WTSP) method for the supervised gene selection and classification scheme. This WTSP method is a generalization of the original TSP method. Different from the TSP method which actually minimizes the sum of misclassification probabilities over two classes, the WTSP minimizes the cost of misclassification or the probability of total misclassification by incorporating the probability of each class in the data. The results obtained from experimental microarray data sets suggest that the WTSP could achieve higher classification accuracy compared to the TSP method. In addition, the WTSP method also simplifies the classification rule by considering the order of the genes in the gene pair. Besides that, the WTSP method possesses the advantages of TSP method such as achieving high classification accuracy with few genes and invariant to the preprocessing.

At this stage, it is difficult to arrive at a conclusion about at what specific conditions the proposed WTSP method can always work significantly better than the original TSP method. The proposed WTSP method aims to handle the problem of the unbalanced sample size in each class. This problem exists in all the 8 data sets we tested. However, for some data sets, the TSP still works comparatively as well as the WTSP. The possible reason is as follows. If the gene pair which achieves the top score in WTSP have $a_{ij} \approx N_1$ and $b_{ij} \approx 0$ (see Eq. (7)), the same gene pair is likely to be chosen in the TSP method. Thus, for those data sets, both methods can achieve a high classification accuracy as shown in the Table 2 for the cases of Leukemia, Colon, Lung and DLBCL. Whereas, if this scenario does not hold, that is, the TSP method works poorly, the proposed WTSP method should improve the performance more significantly.

This is consistent with the results from the GCM, CNS, Prostate and p53 data. Therefore, although it is difficult to pinpoint the exact situation at which the proposed WTSP works significantly better than TSP, a general conclusion is: the WTSP method performs significantly better than TSP method when the sample size in each class is unbalanced and TSP performs poorly.

Future work may include the investigation of other information contained in the gene-gene interaction. The methods based on TSP exploit one pattern of gene-gene interaction, that is, the relative expression levels of the gene pair revert from one class to another class. There may exist other possible gene-gene interaction patterns, such as the coexpression of two genes. By exploring these possible patterns contained in the microarray data, a more accurate classification method could be developed and the functional biological processes may be revealed.

References

1. Dudoit, S., Yang, Y.H., Callow, M.J., Speed, T.P.: Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments. Statistica Sinica 12, 111–139 (2002)
2. Bo, T., Jonassen, I.: New Feature Subset Selection Procedures for Classification of Expression Profiles. Genome Biology 3, research0017.1–research0017.11(2002)
3. Kuo, W.P., et al.: Functional Relationships Between Gene Pairs in Oral Squamous Cell Carcinoma. In: Proceedings of American Medical Informatics Association (AMIA) 2003 Symposium (2003)
4. Hanczar, B., Zucker, J., Henegar, C., Saitta, L.: Feature Construction from Synergic Pairs to Improve Microarray-based Classification. Bioinformatics 23, 2866–2872 (2007)
5. Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., Vert, J.: Classification of Microarray Data Using Gene Networks. BMC Bioinformatics 8 (2007)
6. Karuturi, R.K.M., Vinsensius, B.V.: Friendly Neighbors Method for Unsupervised Determination of Gene Significance in Time-course Microarray Data. In: Proc. of the Fourth IEEE Symposium on Bioinformatics and Bioengineering. IEEE Press, Los Alamitos (2004)
7. Karuturi, R.K.M., Wong, S., Sung, W.K., Miller, L.D.: Differential Friendly Neighbors Algorithm for Differential Relationship Based Gene Selection and Classification using Microarray Data. In: Proc. of the Intl. Conf. on Data Mining (DMIN 2006), USA (2006)
8. Xiong, M., Jin, L., Li, W., Boerwinkle, E.: Computational Methods for Gene Expression Based Tumor Classification. BioTechniques 29, 1264–1270 (2000)
9. Dai, J.J., Lieu, L., Rocke, D.: Dimension Reduction for Classification with Gene Expression Microarray Data. Stat. Appl. Genet. Mol. Biol. 5, 6 (2006)
10. Furey, T., et al.: Support Vector Machine Classification and Validation of Cancer Tissue Samples using Microarray Expression Data. Bioinformatics 16, 906–914 (2000)
11. Duda, R.O., Hart, P.E., Sork, D.G.: Pattern Classification. John Wiley & Sons, New York (2000)
12. Cover, T.M., Hart, P.E.: Nearest Neighbor Pattern Classification. IEEE Trans. Info. Theo. IT 13, 21–27 (1967)

13. Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. J. Amer. Stat. Asso. 97, 77–87 (2002)
14. Tibshirani, R.O., et al.: Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression. Proc. Natl Acad. Sci. 99, 6567–6572 (2002)
15. Tan, A.C., Naiman, D.Q., Xu, L., Winslow, R.L., Geman, D.: Simple Decision Rules for Classifying Human Cancers from Gene Expression Profiles. Bioinformatics 21, 3896–3904 (2005)
16. Xu, L., Geman, D., Winslow, R.: Large-scale Integration of Cancer Microarray Data Identifies a Robust Common Cancer Signature. BMC Bioinformatics 8 (2007)
17. Geman, D., d'Avignon, C., Naiman, D.Q., Winslow, R.: Classifying Gene Expression Profiles from Pairwise mRNA Comparisons. Stat. Appl. Genet. Mol. Biol. 3, 19 (2004)
18. Price, N.D., Trent, J., et al.: Highly Accurate Two-gene Classifier for Differentiating Gastrointestinal Stromal Tumors and Leiomyosarcomas. Proc. Natl Acad. Sci. 104, 3414–3419 (2007)
19. Golub, T.R., et al.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
20. Alon, U., et al.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA 96, 6745–6750 (1998)
21. Gordon, G.J., et al.: Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res. 62, 4963–4967 (2002)
22. Shipp, M.A., et al.: Diffuse large B-cell lymphoma outcome prediction by geneexpression profiling and supervised machine learning. Nat. Med. 8, 68–74 (2002)
23. Ramaswamy, S., et al.: Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA 98, 15149–15154 (2001)
24. Pomeroy, S.L., et al.: Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415, 436–442 (2002)
25. Stuart, R.O., et al.: In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. Proc. Natl Acad. Sci. USA 101, 615–620 (2004)
26. Miller, L.D., et al.: From The Cover: An Expression Signature for p53 Status in Human Breast Cancer Predicts Mutation Status, Transcriptional Effects, and Patient Survival. Proc. Natl. Acad. Sci. USA 102, 13550–13555 (2005)