

L_1 Least Square for Cancer Diagnosis using Gene Expression Data

Xiyi Hang¹, and Fang-Xiang Wu^{2,3*}

¹Department of Electrical and Computer Engineering,
California State University, Northridge, CA 91330, USA

²Department of Mechanical Engineering

³Division of Biomedical Engineering, University of Saskatchewan,
Saskatoon, Saskatchewan, S7N 5A9, Canada

*Corresponding author: Fang-Xiang Wu, Division of Biomedical Engineering
University of Saskatchewan, Saskatoon, Saskatchewan, S7N 5A9,
Canada, E-mail: xhang@csun.edu, faw341@mail.usask.ca

Received March 19, 2009; Accepted April 27, 2009; Published April 27, 2009

Citation: Hang X, Wu FX (2009) L_1 Least Square for Cancer Diagnosis using Gene Expression Data. J Comput Sci Syst Biol 2:167-173. doi:10.4172/jcsb.1000028

Copyright: © 2009 Hang X, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

The performance of most methods for cancer diagnosis using gene expression data greatly depends on careful model selection. Least square for classification has no need of model selection. However, a major drawback prevents it from successful application in microarray data classification: lack of robustness to outliers. In this paper we cast linear regression as a constrained L_1 -norm minimization problem to greatly alleviate its sensitivity to outliers, and hence the name L_1 least square. The numerical experiment shows that L_1 least square can match the best performance achieved by support vector machines (SVMs) with careful model selection.

Keywords: L_1 -norm minimization; Least square regression; Classification; cancer; Gene expression data; Support vector machine

Introduction

DNA microarray technique has the potential to provide a more accurate and objective cancer diagnosis than traditional histopathological approach with its high throughput capability of simultaneously measuring relative expression level of tens of thousands of genes. The success, however, greatly depends upon the supervised learning algorithm selected to classify gene expression data.

Many well-established methods are available for gene expression profile classification. According to Lee et al (2005), they can be classified into four categories: (1) classical methods, such as Fisher's linear discriminant analysis, logistic regression, K -nearest neighbor, and generalized partial least square, (2) classification trees and aggregation methods, such as CART, random forest, bagging and boosting, (3) machine learning methods, such as neural network

and support vector machines (SVMs), and (4) generalized methods, such as flexible discriminant analysis, mixture discriminant analysis, and shrunken centroid method. The performance of many methods, however, relies upon careful choice of model parameters, which can be done via model selection procedure such as cross validation. For example, the model parameters for SVMs include kernel parameters and the penalty parameter C . A recent controversy regarding the performance comparison between SVM and random forest just exemplifies the importance of model selection. The study by Diaz-Uriarte and Alvarez de Andres, (2006) concludes that random forest outperforms SVM, and the conclusion in paper (Stanikov et al., 2008) is totally opposite. The main difference between these two studies is that model selection is carefully designed in the latter study but not in the former study. The incident also shows that model

selection may be the obstacle of the extensive application of SVM in classification of gene expression profile. Since classification performance is a nonconvex function of model parameters, it is usually difficult to find optimal model parameters by model selection.

Least square for classification, on the other hand, has no need of model selection. Consider a general classification problem with N classes. A linear model is built for each class k

$$y_k = \mathbf{w}_k^T \mathbf{x} + w_{k0}, k = 1, 2, \dots, N. \quad (1)$$

The N equations can be grouped into

$$\mathbf{y} = \mathbf{W} \tilde{\mathbf{x}} \quad (2)$$

where $\mathbf{y} = [y_1, y_2, \dots, y_N]^T$, \mathbf{W} is a matrix whose k th row is $[\mathbf{w}_k^T, w_{k0}]$, and $\tilde{\mathbf{x}} = [\mathbf{x}^T, 1]^T$. For a training dataset $\{(\mathbf{x}_i, \mathbf{t}_i), i = 1, 2, \dots, n\}$, where \mathbf{t}_i is 1-of- N binary coding vector of the label of the i th feature \mathbf{x}_i , i.e., a vector containing zeros everywhere except 1 in the k th position, if \mathbf{x}_i belongs to category k . Denote by \mathbf{X} the feature matrix whose k th row is $[\mathbf{x}_k^T, 1]$, and \mathbf{T} the target matrix whose k th row is \mathbf{t}_k^T . The linear regression model in (2) can be fitted simultaneously to each of columns of \mathbf{T} , and the solution is in the form

$$\hat{\mathbf{W}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{T}. \quad (3)$$

- Calculate the fitted output $\hat{\mathbf{y}} = \hat{\mathbf{W}} [\mathbf{x}^T, 1]^T$ (an N vector);
- Label = $\text{argmax}_k \hat{y}(k), k = 1, 2, \dots, N$.

More details can be found in literature (Bishop, 2006; Hastie et al., 2001).

The above approach, however, is very sensitive to outliers, especially for multiclassification ($N \geq 3$). Furthermore, when least square for classification is applied to gene expression data, problems can become more severe due to the curse of dimensionality caused by the great number of genes in each sample.

Inspired by the recent progress in sparse signal recovery via l_1 - norm minimization (Candès et al., 2006, Candès and Tao, 2006; Donoho, 2006), we propose a new approach to overcome the major drawback of least square for classification by casting the linear regression problem as a constrained l_1 - norm minimization problem. The obtained sparse solution is much less sensitive to both outliers and curse of dimensionality. In addition, multiclassification

is realized via one-versus-rest (OVR) and one-versus-one (OVO) approaches which decompose the original multi-category problem into a series of binary problems. The new method is validated by comparing cancer diagnosis performance with SVMs.

Methods

Binary l_1 Least Square

Consider a training dataset $\{(\mathbf{x}_i, y_i); i = 1, \dots, n\}$, $\mathbf{x}_i \in R^d, y_i \in \{-1, +1\}$, where \mathbf{x}_i represents the i th sample, a d -dimensional column vector containing gene expression values with d as the number of genes, and y_i is the label of the i th sample. Two classes are described by a linear model

$$y = [\mathbf{x}^T, 1] \mathbf{w} \quad (4)$$

for any sample \mathbf{x} . Applying the linear model to the training dataset, we have

$$y_i = [\mathbf{x}_i^T, 1] \mathbf{w}, i = 1, 2, \dots, n \quad (5)$$

The n equations can be grouped into

$$\mathbf{y} = \mathbf{X} \mathbf{w} \quad (6)$$

where $\mathbf{y} = [y_1, y_2, \dots, y_n]^T$, and \mathbf{X} is an $n \times (d+1)$ matrix whose i th row is $[\mathbf{x}_i^T, 1]^T$. Since the number of samples are much smaller than the number of genes, i.e., $n \ll d$, the system in (6) is underdetermined. The solution is obtained by casting the original problem as the following constrained l_1 -norm minimization problem

$$\min \|\mathbf{w}\|_1 \text{ subject to } \mathbf{X} \mathbf{w} = \mathbf{y} \quad (7)$$

The above formulation is inspired by the recent progress in compressed sensing (Candès et al., 2006; Candès and Tao, 2006; Donoho, 2006) and basis pursuit denoising (Chen et al., 2005).

There are quite a few solvers available for solving the optimization problem defined in (7), such as MOSEK (Andersen, 2002) PDCCO-CHOL (Saunders, 2002), PDCCO-LSQR (Saunders, 2002), and l_1 -magic (Candès and Romberg, 2006), which all belong to interior-point methods. In this study we choose a solver called SPGL1 (Friedlander and Van den Berg, 2008) for its efficiency in solving large-scale problems. Unlike other methods, SPGL1 solves the optimization problem by converting it into a root finding problem. Please refer to paper (Van den Berg and Friedlander, 2008) for details on the theory of SPGL1.

Denote by $\hat{\mathbf{w}}$ the solution to (7). Then for any sample \mathbf{x} , the label can be simply assigned as $\text{sign}([\mathbf{x}^T, 1] \hat{\mathbf{w}})$.

Multicategory L_1 Least Square: OVR

Consider a multicategory training dataset $\{(\mathbf{x}_i, y_i); i = 1, \dots, n\}$, $\mathbf{x}_i \in R^d, y_i \in \{1, 2, \dots, N\}$, where N is the category number. OVR approach needs to determine for each class a binary classifier to separate it from the remaining classes. The N linear models are defined as

$$D_k(\mathbf{x}) = [\mathbf{x}^T, 1]\mathbf{w}_k, k = 1, 2, \dots, N. \tag{8}$$

For category k , after changing the labels of those samples belonging to k to +1, and others to -1, we apply the linear model to the training dataset

$$\mathbf{y}_k = \mathbf{X}\mathbf{w}_k, k = 1, 2, \dots, N, \tag{9}$$

where \mathbf{y}_k is a label vector containing either +1 or -1. Similarly, the above N underdetermined systems can be solved by the following N constrained L_1 -norm minimization problems

$$\min \|\mathbf{w}_k\|_1 \text{ subject to } \mathbf{X}\mathbf{w}_k = \mathbf{y}_k \tag{10}$$

where $k = 1, 2, \dots, N$.

Denote by $\hat{\mathbf{w}}_k$ the solution to (10). Then for any sample \mathbf{x} , the label can be determined by

$$\arg \max_{k=1,2,\dots,N} D_k(x) = [\mathbf{x}^T, 1]\hat{\mathbf{w}}_k. \tag{11}$$

Multicategory L_1 Least Square: OVO

In OVO approach, a binary classifier is constructed for each pair of classes. The linear model for class i against class j is given by

$$D_{i,j}(\mathbf{x}) = [\mathbf{x}^T, 1]\mathbf{w}_{i,j} \tag{12}$$

For those samples of category i and j , changing their labels to +1 and -1, applying the linear model gives rise to

$$\mathbf{y}_{i,j} = \mathbf{X}_{i,j}\mathbf{w}_{i,j} \tag{13}$$

where $\mathbf{y}_{i,j}$ is a vector containing either +1 or -1, and $\mathbf{X}_{i,j}$ is a matrix whose k th row is $[\mathbf{x}_k^T, 1]^T$ with \mathbf{x}_k belonging to either category i or j . The underdetermined system is solved by

$$\min \|\mathbf{w}_{i,j}\|_1 \text{ subject to } \mathbf{X}_{i,j}\mathbf{w}_{i,j} = \mathbf{y}_{i,j} \tag{14}$$

Since $D_{j,i} = -D_{i,j}$, the number of the classifiers is $\binom{N}{2}$, i.e., $N(N-1)/2$, compared to N in OVR approach.

Denote by $\hat{\mathbf{w}}_{i,j}$ the solution to (14). For any sample \mathbf{x} , we calculate

$$D_i(x) = \sum_{j=1, j \neq i}^N \text{sign}(D_{i,j}(x)) \tag{15}$$

with $D_{i,j}(\mathbf{x}) = \sum_{j=1, j \neq i}^N \mathbf{X}_{i,j}\hat{\mathbf{w}}_{ij}$. The label of \mathbf{x} is determined

by

$$\arg \max_{i=1,2,\dots,N} D_i(x) \tag{16}$$

Numerical Experiment

Numerical experiment is carefully designed to validate the cancer diagnosis performance of 11 least square using gene expression data. The performance metric is classification accuracy obtained by 10-fold stratified cross validation. MATLAB R14 is used to implement the new method. The results are compared with binary SVM (Vapnik, 1998) and some popular variants of multicategory SVMs including OVR-SVM (Kressel, 1999), OVO-SVM (Kressel, 1999), DAGSVM (Platt et al., 2000), method by Weston and Watkins (WW) (Weston and Watkins, 1999), and method by Crammer and Singer, (2000).

The results of SVMs are obtained from GEMS (Gene Expression Model Selector), which is software with graphic user interface for classification of gene expression data. It is freely available at <http://www.gems-system.org/>. GEMS is used by Stanikov et al., (2005) for the comprehensive study of the performance of multiple classifiers on gene expression cancer diagnosis. As for model selection, polynomial kernels are used with orders $p = \{1, 2, 3\}$, and the penalty parameter $C = \{10^{-3+0.5n}, n = 0, 1, \dots, 6\}$.

Six datasets are used in the experiment, which are among eleven datasets used in reference (Stanikov et al., 2005). They are available on the website of GEMS in the format of both GEMS and MATLAB mat file. For easy comparison and reference, we adopt the names used in reference (Stanikov et al., 2005). The information about the six datasets is summarized below.

- DLBCL (Shipp et al., 2002): The binary dataset comes from a study of gene expression of two lymphomas: diffuse large B-cell lymphomas and follicular lymphomas. Each sample contains 5469 genes. The sample number is 77.
- Prostate_Tumor (Singh et al., 2002): The binary dataset contains gene expression data of prostate tumor and normal tissues. There are 10509 genes in each sample, and 102 samples.
- 9_Tumors (Staunton et al., 2001): The dataset comes from a study of 9 human tumor types: NSCLC, colon, breast, ovary, leukaemia, renal, Melanoma, prostate, and CNS. There are 60 samples, each of which contains 5726 genes.
- 11_Tumors (Su et al., 2001): The dataset includes 174

samples of gene expression data of 11 various human tumor types: ovary, bladder/ureter, breast, colorectal, gastroesophagus, kidney, liver, prostate, pancreas, lung adeno, and lung squamous. The number of genes is 12533.

- Brain_Tumor1 (Pomeroy et al., 2002): The dataset comes from a study of 5 human brain tumor types: medulloblastoma, malignant glioma, AT/RT, normal cerebellum, and PNET, including 90 samples. Each sample has 5920 genes.
- Brain_Tumor2 (Nutt et al., 2003): There are 4 types of malignant glioma in this dataset: classic glioblastomas, classic anaplastic oligodendrogliomas, non-classic glioblastomas, and non-classic anaplastic oligodendrogliomas. The dataset has 50 samples, and the number of genes is 10367.

All the datasets are normalized by rescaling the gene expression values to be between 0 and 1.

Two methods are used in this experiment to study gene selection's impact on classification performance: Kruskal-Wallis non-parametric one-way ANOVA (KW) (Gibbons, 2003), and the ratio of between classes to within class sums of square (BW) (Dudoit et al., 2002).

Results

Classification without Gene Selection

Table 1 shows the classification accuracy values obtained by 10-fold stratified cross validation for both l_1 least square and SVMs. The results of SVMs are slightly different from what is reported by Stanikov et al., (2005) where the five datasets are also used. A possible explanation is that the distribution for cross validation in our study is different from that in paper (Stanikov et al., 2005).

For binary datasets Prostate_Tumor and DLBCL, the performance of l_1 least square is slightly below that of SVMs. Note that the results of SVMs are obtained by careful model selection using cross validation, while our method does not need model selection, and is totally automatic. In addition, just like SVM, when applied to binary datasets, the multicategory classifiers of l_1 least square are equivalent to binary classifier for both OVO and OVR approaches.

When applied to classification of multicategory datasets, OVR- l_1 least square can closely match the best performance achieved by SVMs. For both SVM and l_1 least square, OVO approach performs much worse than OVR approach for classifying 9 Tumors dataset.

| Methods | | Prostate Tumor | DLBCL | 9 Tumors | 11 Tumors | Brain Tumor1 | Brain Tumor2 |
|-----------|--------|----------------|--------|----------|-----------|--------------|--------------|
| SVM | Binary | 93.27% | 97.32% | N/A | N/A | N/A | N/A |
| | OVR | 93.27% | 97.32% | 67.06% | 94.99% | 90% | 75.5% |
| | OVO | 93.27% | 97.32% | 54.63% | 90.22% | 90% | 73.83% |
| | DAGSVM | 93.27% | 97.32% | 54.63% | 90.22% | 90% | 73.83% |
| | WW | 93.27% | 97.32% | 68.17% | 94.31% | 90% | 77.17% |
| CS | 93.27% | 97.32% | 68.17% | 94.31% | 90% | 75.5% | |
| l_1 LRC | Binary | 91.36% | 96.07% | N/A | N/A | N/A | N/A |
| | OVR | 91.36% | 96.07% | 72.21% | 96.63% | 90% | 76.67% |
| | OVO | 91.36% | 96.07% | 55.33% | 91.93% | 90% | 77.00% |

Table 1: Performance without gene selection.

| Methods | | Prostate Tumor | DLBCL | 9 Tumors | 11 Tumors | Brain Tumor1 | Brain Tumor2 |
|-----------|----------|----------------|--------|----------|-----------|--------------|--------------|
| SVM | Accuracy | 94.27% | 98.75% | 72.89% | 96.66% | 90% | 82.83% |
| | Variant | OVO | OVO | CS | OVR | WW | OVR |
| | GS | KW 1000 | KW 500 | BW 3000 | KW 1000 | NG | KW 500 |
| OVR | Accuracy | 94.18% | 98.75% | 75.69% | 96.66% | 90% | 78.33% |
| l_1 LRC | GS | BW 3050 | BW 500 | KW 1060 | KW 2000 | NG | BW 9000 |

Table 2: Performance with gene selection.

Classification with Gene Selection

Table 2 shows the best performance achieved by OVR- l_1 least square and SVMs when gene selection methods KW and BW are used. The results show that both l_1 least square and SVMs perform slightly better compared with the performance without gene selection reported in Table 1. The improvement ranges from 0 to 9% for SVMs, while only from 0 to 3.48% for OVR- l_1 least square. Again, the performance of OVR- l_1 least square is comparable to SVMs.

Discussion

The success of l_1 least square may lie in its sparse linear model coefficient vector obtained from l_1 - norm minimization. Figure 1 shows the model coefficient vector w which is the solution of l_1 least square for classifying binary dataset DLBCL. The sparsity suggests that those genes with greater absolute coefficients could have played more important roles in classification. As a result, the classification performance does not depend on all the genes, especially those with very small absolute coefficients. The sparsity has the potential to greatly alleviate curse of dimensionality and increase the robustness to outliers.

Another implication of sparsity is that those genes with larger absolute coefficients may correspond to biological markers. Hence, sparsity could be also used for gene selection. We did a small experiment to verify this possibility. The binary dataset DLBCL is used to fit l_1 least square

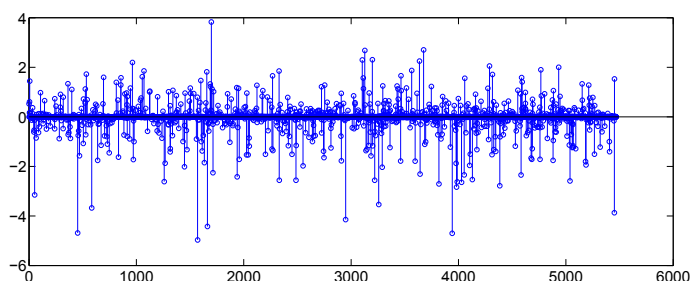


Figure 1: The sparse coefficient vector.

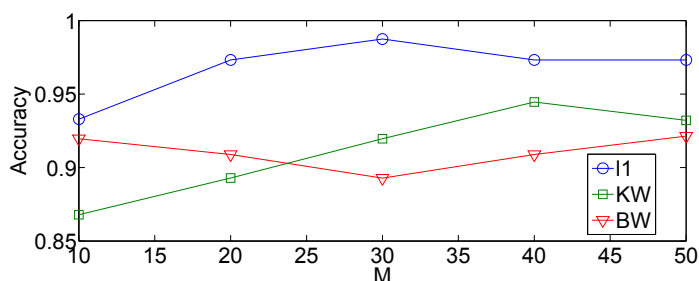


Figure 2: The performance of three gene Selection methods.

model. Gene selection is done by choosing M genes with M largest absolute coefficients. Binary SVM is used to classify the gene-selected data. The results are compared with KW and BW methods for gene selection. Figure 2 shows the performance of the three gene selection methods for $M = 10, 20, 30, 40,$ and $50,$ respectively. The new method significantly outperforms both KW and BW methods when a small number of genes are selected.

The above gene selection approach is in spirit similar to lasso (Tibshirani, 1996) formulated as follows

$$\min \| \mathbf{X}\mathbf{w} - \mathbf{y} \|_2^2 \quad \text{subject to } \| \mathbf{w} \|_1 \leq t \tag{17}$$

where $\mathbf{X}, \mathbf{w},$ and \mathbf{y} follow the definitions given in section 2.1 for binary l_1 least square, and t is the model parameter for lasso. In addition, lasso can also be used in classification by replacing (7) with (17) for binary case, (10) with

$$\min \| \mathbf{X}\mathbf{w}_k - \mathbf{y}_k \|_2^2 \quad \text{subject to } \| \mathbf{w}_k \|_1 \leq t_k \tag{18}$$

for multi-category OVR approach, and (14) with

$$\min \| \mathbf{X}\mathbf{w}_{i,j} - \mathbf{y}_{i,j} \|_2^2 \quad \text{subject to } \| \mathbf{w}_{i,j} \|_1 \leq t_{i,j} \tag{19}$$

for multi-category OVO approach.

Similarly, we can also replace l_1 least square regression with Dantzig selector (Candès and Tao, 2007), which is given below for binary classification

$$\min \| \mathbf{w} \| \quad \text{subject to } \| \mathbf{X}^T (\mathbf{y} - \mathbf{X}\mathbf{w}) \|_\infty \leq (1 + t^{-1}) \sqrt{2 \log d} \sigma \tag{20}$$

where t is model parameter, and σ is the noise standard deviation. Dantzig selector for multcategory classification can be similarly defined.

Both lasso and Dantzig selector for classification, however, still need to select optimized model parameters by model selection procedure, such as cross validation.

Conclusion

In this paper, we have described a specialized regression method for cancer diagnosis using expression data. The new approach, called l_1 least square, casts linear regression as a constrained l_1 -norm minimization problem to overcome the major drawback of least square for classification: lack of robustness to outliers. Besides binary classifier, multcategory l_1 least square including OVO and OVR approaches are also proposed.

Numerical experiment shows that OVR- l_1 least square can match the best performance achieved by SVMs with careful model selection. The main advantage of l_1 least

square over other methods including SVMs is that it has no need of model selection. As a result, the method based on l_1 least square is totally automatic. l_1 least square also has the potential to be used for gene selection.

The l_1 least square classifier may become a promising automatic cancer diagnosis tool by consistently distinguishing gene profile classes. Those genes with great absolute regression coefficients may serve as biological marker candidates for further investigation.

References

1. Bishop CM: *Pattern recognition and machine learning*. New York: Springer; 2006. » [CrossRef](#) » [Google Scholar](#)
2. Candès E, Romberg J, Tao T (2006) Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans Inform Theory* 52: 489-509. » [CrossRef](#) » [Google Scholar](#)
3. Candès EJ, Tao T (2006) Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. on Information Theory* 52: 5406 - 5425.
4. Candès EJ, Romberg J (2006) l_1 -magic: A Collection of MATLAB Routines for Solving the Convex Optimization Programs Central to Compressive Sampling [Online]. Available: www.acm.caltech.edu/l1magic/
5. Candès E, Tao T (2007) The Dantzig selector: Statistical estimation when p is much larger than n . *Ann Statist* 35: 2313-2351. » [CrossRef](#) » [Google Scholar](#)
6. Chen SS, Donoho DL, Saunders MA (2001) Atomic decomposition by basis pursuit. *SIAM Rev* 43: 129-159. » [CrossRef](#) » [Google Scholar](#)
7. Crammer K, Singer Y (2000) On the learnability and design of output codes for multiclass problems. *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*. Stanford University Palo Alto CA June 28–July 1.
8. Diaz-Uriarte R, Alvarez de Andres S (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinform* 7: 3. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
9. Donoho DL (2006) Compressed sensing. *IEEE Trans Inform Theory* 52: 1289-1306. » [CrossRef](#) » [Google Scholar](#)
10. Dudoit S, Fridlyand J, Speed TP (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc* 97: 77-87. » [CrossRef](#) » [Google Scholar](#)
11. Friedlander M, Van den Berg E (2008) SPGL1, a solver for large scale sparse reconstruction. [Online] Available: <http://www.cs.ubc.ca/labs/scl/spgl1/>
12. Gibbons JD: *Nonparametric Statistical Inference*, 4th edition, CRC, 2003.
13. Hastie T, Tibshirani R, Friedman J (2001) *The elements of statistical learning*. New York: Springer. » [CrossRef](#) » [Google Scholar](#)
14. Kressel U (1999) Pairwise classification and support vector machines. In *Advances in Kernel Methods: Support Vector Learning*, (Chapter 15.) Cambridge, MA: MIT Press.
15. Lee JW, Lee JB, Park M, Song SH (2005) An extensive comparison of recent classification tools applied to microarray data. *Comput Stat Data Anal* 48: 869-885. » [CrossRef](#) » [Google Scholar](#)
16. Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, et al. (2003) Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res* 63: 1602-1607. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
17. Platt JC, Cristianini N, Shawe-Taylor J (2000) Large margin DAGS for multiclass classification. In *Advances in Neural Information Processing Systems* 12. MIT Press. » [Google Scholar](#)
18. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, et al. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415: 436-442. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
19. Saunders M (2002) PDCO: Primal-Dual Interior Method for Convex Objectives [Online]. Available: <http://www.stanford.edu/group/SOL/software/pdco.html>.
20. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, et al. (2002) Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. *Nat Med* 8: 68-74. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
21. Statnikov A, Wang L, Aliferis CF (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics* 9: 319. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)

22. Staunton JE, Slonim DK, Collier HA, Tamayo P, Angelo MJ, et al. (2001) Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci USA* 98: 10787-10792. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
23. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* 21: 631-643. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
24. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, et al. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 203-209. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
25. Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, et al. (2001) Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res* 61: 7388-7393. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
26. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J Roy Statist Soc ser B* 58: 267-288. » [CrossRef](#) » [Google Scholar](#)
27. The MOSEK Optimization Tools Version 2.5. User's Manual and Reference 2002 [Online]. Available: www.mosek.com.
28. Van den Berg E, Friedlander M (2008) Probing the Pareto frontier for basis pursuit solution. Technical Report 2008, Department of Computer Science, University of British Columbia.
29. Vapnik VN: *Statistical learning theory*. New York: Wiley; 1998.
30. Weston J, Watkins C (1999) Support vector machines for multi-class pattern recognition. In *Proceedings of the Seventh European Symposium On Artificial Neural Networks (ESANN 99) Bruges April 21-23*.