Journal of Data Science **2**(2004), 165-175

Identify Breast Cancer Subtypes by Gene Expression Profiles

Grace S. Shieh[1] Chy-Huei Bai[1] and Chih Lee[2]
[1]*Academia Sinica and* [2]*National Taiwan University*

Abstract: Support vector machines (SVMs), with linear, polynomial and radial kernels, were applied to classify subtypes of breast cancer by gene expression profiles of tissues samples. Using the top 500 genes ranked by between-group to within-group sum of squares, SVMs with linear kernel had an average accuracy rate about 97% when applied to a balanced dataset; this accuracy rate was significantly higher than that of the original data. After imputation, the smallest subsample of the balanced dataset was comparable to the other subsamples' (containing more than 10 samples). In biomedical sciences, it is of interest to identify genes that can be used to classify subtypes of breast cancer well. Using SVMs, we identified 500 genes and looked up the functions of 297 genes from databases. Furthermore, about 65% of these 297 genes were known to be related to breast cancer, and this confirms the consistency of our results with existing biomedical knowledge. Those 203 genes may also be investigated further to see if they are involved in breast cancer; any novel findings will be important.

Key words: Classification, microarray gene expression data, support vector machines, tumor.

1. Introduction

Breast cancer is the most common malignancy among women. The clinical symptoms of breast cancer patients are heterogeneous and they depend primarily on whether the disease is metastasized or not. Traditionally, a thorough evaluation of a breast cancer patient includes an evaluation of the extent of disease and the assessment of the grade of histology type, tumor size, auxiliary lymph nodes status, hormone receptor and ErbB2 receptor statuses. However, tumors with identical traditional diagnosis or histopathology may progress differently, respond to therapy differently, and may result in different disease outcomes. Genome-wide analyses of gene expression profiles suggested that different subclasses of tumors, for instance Leukemia (B-cell versus T-cell) and breast cancer (ER+ versus ER−), correspond to distinct gene expression patterns (Golub *et al.* 1999 and Gruvberger *et al.*, 2001, respectively). Early prediction on tumor

subclasses can be incorporated by physicians to make therapeutical decision, and to improve the prognosis of carcinoma patients (Van De Vijver *et al.*, 2002). Thus microarray data analysis may serve as a new and comprehensive tool to guide therapeutic interventions in the near future.

Sorlie *et al.* (2001) clustered 85 gene expression profiles into six subclasses according to the histologic and clinical outcomes. Recent researches suggested that classification of tumors based on gene expression patterns could be used as a prognostic maker (Golub *et al.*, 1999; Alizadeh *et al.*, 2000; Perou *et al.*, 2000 and Sorlie *et al.*, 2001). Furthermore, Sorlie *et al.* (2001) found that patients belonging to different subclasses had significantly different outcomes from a survival analysis. These suggested that expression levels of these genes can be used to classify tumor subtypes well may provide information on cancer patients' prognosis, and thus these genes are of interest. Furthermore, using microarray expression levels of these genes solely (without other types of clinical data) can help predict breast cancer subtypes and thus this may be a diagnosis tool in the future.

In a pilot study, support vector machines (SVMs) with linear, polynomial and radial kernels (Vapnik, 1998), Fisher's discrimination analysis (both FLDA and QLDA, Johnson and Wichern, 1992) and CRUISE (Kim and Loh, 2001) were applied to classify the 85 profiles into six subclasses. Each profile contained 6,228 gene expression levels. The preliminary results can be downloaded on the web site `http://stat.sinica.edu.tw/~gshieh/bcancer.ppt`. Among the three classifiers, SVMs performed the best in terms of accuracy rate. Thus, we applied SVMs to classify gene expression profiles of the samples tissues in this article.

We investigated how to classify these 85 gene expression profiles into six subclass well with minimum number of genes. An introduction of the microarray gene expression data and data pre-processing are in Section 2. The methodology of SVMs, criterion of feature selection and two methods to balance the subsample sizes are given in Section 3. Section 4 reports the classification experiments and results of SVMs with three types of kernels, using top 10 to 1,000 ranked genes. We close with some remarks in Section 5.

2. cDNA Microarray Gene Expression Data

In this section, we first provide an introduction of the microarray dataset used in Section 2.1. Data pre-processing and imputation are in Section 2.2.

2.1. Dataset

Gene expression profiles form 78 breast carcinomas (71 ductal, 5 lobular, and

2 ductal carcinomas *in situ*), 3 fibroadenomas and 4 normal breast samples were used in this study. The detailed list of all 85 samples with at least 9,216 genes and their clinical data have been published on the PNAS web site, `www.pnas.org`. Eighty-five profiles, originally derived from Stanford Microarray Database, were downloaded from the web site `http://genome-www4.stanford.edu/MicroArray/SMD/`. All experiments and the production of microarrays were performed as described in Perou *et al.* (2000). Sorlie *et al.* (2001) incorporated clinical outcomes (survival, survival time, relapse, tumor category, node status, tumor grade and metastasis) and microarray gene expression data of 85 samples. applied significance analysis of microarray (SAM) method (Tusher *et al.*, 2001) to cluster the 85 microarray gene expression profiles, using 427 genes, into six subclasses. These subclasses are basal-like (14 samples), ERBB2+ (11 samples), normal basal-like (13 samples), luminal subtype A (32 samples), subtype B (5 samples), and subtype C (10 samples), respectively.

In this article, we used the six subclasses as correct subtypes of the 85 profiles and studied performances of SVMs with different types of kernels. Thus Sorlie *et al.* (2001) tackled a unsupervised (clustering) problem while we studied a supervised problem. Special emphasis was given to finding genes that were top-ranked by the BSS/WSS criterion since they were potential markers for gene expression profiles to identify breast cancer subtypes.

2.2. Data preprocessing and imputation

The downloaded data were filtered and normalized by ScanAlyze version 2.5 and Genepix Pro 5.0 (Axon Instruments, Foster City, CA). For a given gene i with mRNA from sample j, denoted by x_{i_j}, the ratio of RD to GDN was used as its gene expression level, where RD is the difference between the gene's red channel intensity and its background intensity. Likewise, GDN is the normalized difference between the gene's green channel intensity and its background such that $\log(\mathrm{RD}/\mathrm{GDN}) = 0$ for any empty spot on an array. Although each array consisted of at least 9,216 genes, we used only 6,228 genes which were common in all 85 gene expression profiles. For a given gene which had replications in one array, its average was used to represent its gene expression level. We imputed those missing gene expression levels by the average of other genes' expression levels present in the same subclass.

3. Methodology

In this section, we concisely introduce SVMs in Section 3.1. Criterion of variables (feature) selection and two methods to balance the subsample sizes are in Sections 3.2 and 3.3, respectively.

3.1 SVMs

SVMs were originated from the statistical learning theory of Vapnik and co-workers in the 1970s (Vapnik, 1998). SVMs have been popular since the 1990s due to the advancement of modern computing. When used in classification, SVMs separate binary labeled training data by constructing a hyperplane, which separates class members from non-members. A maximum margin hyperplane is the hyperplane that has maximal distance from members to non-members. When the data are not linear-separable, SVMs map the data into a higher dimensional space (called a feature space), and define a separating hyperplane there. The kernels of the SVMs automatically realize a non-linear mapping to a feature space. The hyperplane found by an SVM in the feature space corresponds to a decision boundary in the input space (Cristianini and Shawe-Taylor, 2000).

SVMs with three types of kernels, linear, polynomial and radial basis were applied, and they are denoted by SVM(L), SVM(P) and SVM(R), respectively. The kernel measures the similarity between sample profiles $\mathbf{X_i}$ and $\mathbf{X_j}$, where $1 \leq i, j \leq 85$ and each profile contains gene expression levels of the top-k genes used and the sample's subclass label. Specifically, $\mathbf{X_i} = (X_{i1}, \cdots, X_{ik})$ denote the expression levels of top-k genes ranked by the feature selection criterion in Section 3.2, so it is a vector of k components. The three types of kernels of SVMs are measures of similarity between two profiles. These kernels assume the following forms: $k(\mathbf{X_i}, \mathbf{X_j}) = \mathbf{X_i} \cdot \mathbf{X_j}$, $k(\mathbf{X_i}, \mathbf{X_j}) = (\mathbf{X_i} \cdot \mathbf{X_j} + 1)^2$ and $k(\mathbf{X_i}, \mathbf{X_j}) = \exp(-\|\mathbf{X_i} - \mathbf{X_j}\|)^2/2\sigma^2)$, where $\|\mathbf{X_i}\| = \sum_{l=1}^{k} X_{il}^2$ and σ is the standard deviation of a Gaussian distribution. The parameter σ in the radial kernel is a scaling parameter that penalizes the dissimilarity, namely a small value of σ gives a big dissimilarity value and vice versa. For simplicity, all default values of parameters in the SVM(P) and SVM(R) were used; $p = 2$, $C = 1$ and $2\sigma^2 =$ the number of subclasses.

Since there are 6 subclasses of breast cancer in the dataset, we applied a multiclass SVM algorithm (Hsu and Lin, 2002). Basically, there are two methods to extend a 2-class SVMs algorithm to a multiclass one, namely one-against-one and one-against-all (Rifkin and Klautau, 2004). The algorithm we applied incorporated the one-against-one method. In our case, for a given profile there were 15 2-class subclass votings and the majority vote classified the profile into the corresponding subclass.

3.2 Criterion for feature selection

We used the feature (gene) selection criterion in Dutoit *et al.* (2002), namely the ratio of genes' between-group to within-group sum of squares (*BSS/WSS*). This ratio compares 'the distance of the center of each class to the over-all center'

to 'the distance of each gene to its class center'. The formula of *BSS/WSS* for a given gene j has the form:

$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{.j})^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2},$$

where y_i denotes the subclass label of gene i, $\bar{x}_{.j}$ denotes the average expression level of gene j across all samples and $\bar{x}_{kj}$ denotes the average expression level of gene j belonging to subclass k and $k = 1, \cdots, 6$. Performances of SVMs with these three types of kernels using top-p genes ranked by BSS/WSS, where p ranges from $10, 20, \cdots$, to 100 and from $100, 200, \cdots$ to to $1,000$ are compared in Table 1 in Section 4.

3.3 Balancing sample sizes of subclasses

Using the preprocessed data, even with the top-1000 genes ranked by the ratio of BSS to WSS, the prediction accuracy rate of SVM(L) was limited to 86%. This is due to a very small sample size of subclass 5 (luminal subtype B) whose accuracy rate was about 63%. The sample size of subclass 5 was small compared to the others (having at least 10 profiles). This unbalanced subsample problem seemed to confine performance of SVMs since the accuracy rates of SVMs with all types of kernels applied to the unbalanced dataset was all limited to 86%. Thus we utilized two methods to balance the very small subsample size of subclass 5.

Method one duplicated all items in subclass 5 (luminal subtype C) three times so that its *balanced* sample size (15) was compatible to the others'. Method two assumed that the logarithm gene expression levels in each subclass followed a normal distribution $N(\mu_i, \sigma_i^2)$, where μ_i and σ_i^2 would be estimated from all data in subclass i. Then we randomly and independently sampled 10 items from $N(\hat{\mu}_5, \hat{\sigma}_5^2)$ and added them to subclass 5. Finally, Method one was adopted due to high accuracy rate. SVMs was then applied to the balanced data.

4. Experiments and Results

Details of the cross-validation classification experiments are described in Section 4.1. Results of SVMs with three types of kernels, using top 10 to 1,000 ranked genes are in Section 4.2; cellular roles of 297 (from the top-500) genes have been identified therein.

4.1 Experiments

Supervised learning and testing were performed on the data set by randomly sampling data from each subclass with the ratio of 2:1. Let n_k denote the number

of genes in subclass k, $l_k = [(2n_k)/3]$ the number of genes in the learning set of subclass k, where $[a]$ is the Gauss integer of a, Then the number of genes in the test set of subclass k t_k, is equal to $n_k - l_k$. Finally, let T be the total number of genes in all test sets. The prediction accuracy rate is computed by the ratio of the number of genes been predicted correctly to T in each experiment.

This experiment was repeated 150 times randomly chosen from a complete 3-fold cross-validation experiment, and an average prediction accuracy rate was computed. Similar to the prediction accuracy rate, the subclass prediction accuracy rates were also calculated by the ratio of 'the number of genes been predicted correctly in subclass k' to t_k. Accuracy rates were computed by a C++ program.

4.2 Results

Let BSS/WSS denote the ratio of sum of squares of between-class to within-class. Table 1 shows the accuracy of SVMs applied to classify these 85 profiles using expression levels of the top 10 genes to those of the top 1,000 genes ranked by BSS/WSS.

Table 1: Accuracy rates of SVMs with three types of kernels, applying to data balanced by Method One, using top-10 to top-1000 genes ranked by BSS/WSS

# genes used	SVM		
	Linear	Order 2 Polynomial	Radial
10	82.4	81.9	85.6
20	89.1	89.2	93.6
30	93.3	92.6	95.3
40	94.6	94.5	96.2
50	92.5	92.7	93.8
60	94.3	94.8	95.6
100	97.0	96.7	95.1
200	96.7	94.8	94.1
300	96.3	94.2	94.2
400	95.8	93.5	94.2
500	96.7	93.7	93.6
600	96.8	93.9	94.3
700	95.5	92.4	93.3
800	96.6	93.7	93.2
900	96.1	93.0	92.5
1,000	95.6	92.5	91.6

Table 2: Accuracy rates of SVMs with three types of kernels, applying to data balanced by Method Two, using top-100 to top-1000 genes ranked by BSS/WSS

# genes used	SVM		
	Linear	order 2 Polynomial	Radial
100	96.5	95.2	95.7
200	96.5	94.1	94.7
300	96.9	93.1	93.9
400	96.2	92.3	93.7
500	95.8	92.4	94.2
600	95.7	91.9	94.0
700	95.3	92.0	93.9
800	95.4	92.4	93.5
900	94.5	91.7	92.8
1000	94.7	91.7	91.1

The results of SVMs applied to data imputed by Methods One and two in Section 3.3 are in Tables 1 and 2, respectively. The average accuracy rate increased as the number of genes (features) increased, and the accuracy rate did not increase essentially after the top 500 genes used. From the cross-validation experiments, we obtained that SVMs with linear, polynomial (of order 2) and radial kernels were about 96.7, 93.7 and 93.6, respectively. Using these top-500 genes ranked by BSS/WSS, SVM(L) performed the best among the three types of kernels compared.

SVMs applied to each subclass of these 85 profiles from the original data (without any balancing procedures), resulted in quite different accuracy rates. Among them, the accuracy rate of SVMs applying to subclass luminal C was the worst. Apparently this was due to its subsample size being too small (5) compared to the rest (at least 10). Furthermore, the accuracy rate of SVMs applying to subclass ERBB2+ was the second to the worst, and again this poor rate was likely due to its small sample size (10) compared to the others (about 15). However, even we increased the subsample sizes of subclasses ERBB2+ after balancing subclass luminal C, the average accuracy rate did not increase much (still about 97%). Thus for simplicity, we only increased the sample size of subclass luminal C by Method one in Section 3.3. Using the original (unbalanced) dataset, the average accuracy rate of SVMs was limited to 86% even with the top 1,000 ranked genes used. With the balanced dataset, the accuracy rate of SVMs was 97% using the top 500 ranked genes. This showed that balanced subsample size was a key for SVMs to work well for a multiclass classification problem.

After those subclass sample sizes balanced, the average classification accuracy rate increased from 89% (unbalanced subsample sizes) to 97% using the top-500 genes ranked by the BSS/WSS ratios. The details were shown in Table 3.

Table 3: Subclass accuracy rates of SVMs using the top-500 genes ranked by BSS/WSS

	SVM(L)[a]	SVM(P)[b]	SVM(R)[c]
Overall average accuracy rate	96.7	93.7	93.6
Subclasses accuracy rates			
Basal-like	100.0	97.8	100.0
ERBB2+	90.0	84.9	84.7
Normal basal-like	97.2	89.3	90.8
Luminal subtype A	97.9	95.2	99.9
Luminal subtype B	98.0	97.2	97.2
Luminal subtype C	92.0	92.0	70.7

[a]denotes SVMs with linear kernel, [b]denotes SVMs with polynomial kernel (of order 2), [c]denotes SVMs with radial kernel.

As suggested in Sorlie *et al.* (2001) that those genes that could classify subtypes of breast cancer well might serve as prognosis markers and thus their functions are of interest. Functions of these top-500 ranked genes were checked by the Humam Genome Index (H.G.I.) at TIGR web site. The cellular functions of 297 genes were identified, while the other 203 genes could not be identified.

Table 4: Cellular roles of the 297 genes and their percentages

	% among the 297 genes
Cell division	6.1
Cell signaling/cell communication	23.2
Cell structure/motility	11.4
Cell/organism defense	5.7
Gene/protein expression/RNA synthesis	19.2
Metabolism	22.6
Unclassified	11.8

Table 4 shows the percentages of cellular role categories of the 297 genes based on the categorization of H.G.I. Certain functions, cell signaling/communication, metabolism and gene/protein expression, are known to be related to breast carcinoma. Genes involved with these functions are about 23.2%, 22.6% and 19.2%,

respectively among the 297 genes. Overall, the results are consistent with existing knowledge in medical sciences. These 500 genes may be designed into a diagnosis array for breast carcinoma subtypes that may help predict distinct prognosis statuses.

5. Conclusion

SVMs with three types of kernels were applied to 85 gene expression profiles from Sorlie *et al.* (2001) for classification of breast cancer subtypes. Using the top 500 ranked genes, the average accuracy rate of SVMs applied to the balanced dataset was about 97%, compared to 86% using the original unbalanced dataset. Thus balanced subsample size seemed to be a key for SVMs to work well in this multiclass classification problem. Further study on this issue is of interest. Among the top 500 ranked genes, functions of 297 were identified, and about 65% of these functions were known to be important to breast cancer. These 500 genes may be investigated further by biomedical experiments; any novel findings will be important in medical sciences.

Acknowledgement

We thank Dr. Yuan-Chin Chang for helpful discussions, an anonymous referee and the editor for constructive comments. This research was supported in part by the National Science Council of Taiwan, R.O.C. through grants 91-3112-P001-037 and 91-2118-M001-018 to G.S.S.

References

Alizadeh, A. A., *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees.* Chapman and Hall

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines.* Cambridge University Press.

Dutoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *J. Amer. Statis. Assoc.* **97**, 77-86.

Golub, T. R., Slomin, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P. Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537.

Gruvberger, S., Ringner, M., Chen, Y., Panavally, S., Saal, L. H., Borg, A., Ferno, M., Peterson, C. and Meltzer, P. (2001). Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.* **61**, 5979-5984.

Hsu, C.-W. and Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Networks* **13**, 415-425.

Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis,* 3rd Edit. Prentice Hall.

Kim, H. and Loh, W.-Y. (2001). Classification trees with unbiased multiway splits. *J. Amer. Statis. Assoc.* **96**, 589-604.

Perou, C. M., *et al.* (2000). Molecular portraits of human breast tumors. *Nature* **406**, 747-52.

Rifkin, R. and Klautau, A. (2004). In defense of one-vs-all classification. *J. Machine Learning Research* **5**, 101-141.

Shieh, G. S., Jiang, Y.-C. and Shih, Y.-S. (2003). Comparison of SVMs to some classifiers using gene expression data. Submitted to *Comput. Statist. Data Analy.*.

Sorlie, T., *et al.* (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, *Proc. Natl. Acad. Sci. USA*. **98** 10869-74.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significant analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.* **98**, 5116-5121.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999-2009 .

Vapnik, V. (1998). *Statistical Learning Theory.* Wiley.

Received Febuary 9, 2004; accepted March 6, 2004.

Grace S. Shieh
Institute of Statistical Science
Academia Sinica, Taipei 115
TAIWAN, R.O.C.
gshieh@stat.sinica.edu.tw

Chy-Huei Bai
Institute of Biomedical Sciences
Academia Sinica, Taipei 115
Taiwan, R.O.C.

baich@gate.sinica.edu.tw

Chi Lee
Department of Computer Science and Information Engineering
National Taiwan University, Taipei 106
Taiwan, R.O.C.