# A Strategy of finding the optimal number of features on gene expression data

Alok Sharma, Chuan Hock Koh, Seiya Imoto, Satoru Miyano

Feature selection is considered to be an important step in the analysis of transcriptomes or gene expression data. Carrying out feature selection reduces the curse of dimensionality problem and improves the interpretability of the problem. Numerous feature selection methods have been proposed in literature and these methods rank the genes in order of their relative importance. However, most of these methods determine the number of genes to be used in an arbitrarily or heuristically fashion. In this Letter, we propose a theoretical way to determine the optimal number of genes to be selected for a given task. We applied this proposed strategy on a number of gene expression datasets and obtained promising results.

*Introduction:* Dimensionality reduction techniques are applied to high dimensional problems for reducing computational complexity and improving generalization performance. Various dimensionality reduction techniques can be grouped into two categories, namely, feature extraction and feature selection. In feature extraction, feature vectors are transformed into a parsimonious data space using linear or non-linear combination of feature vectors; and, in feature selection, only some important features or attributes are retained and the remaining features are discarded. Feature selection methods play a crucial role

in the identification of important genes responsible for characterizing heterogeneity of human cancers.

Numerous feature selection methods have been proposed in the literature [1],[2],[3],[4]. A comprehensive study can be found in Ref. [5]. These methods explore the significance of genes and rank them based on a certain feature score. Then, top $h$ genes are selected for downstream application such as classification or clustering. Typically, the value of $h$ is selected arbitrarily which could lead to suboptimal performance. It has also been observed in many situations where the chosen $h$ is too large and a much lower $h$ would achieve similar or even better performance. In this paper, we propose a theoretically founded strategy to select the optimal $h$ that ensures minimum error rate with currently available training data. Using several publicly available gene expression datasets, we demonstrate the utility and performance of this strategy.

*Proposed strategy:* The mathematical notations used in this Letter are defined as follows. Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a set of *n* training vectors in a *d*-dimensional feature space. Let $\Omega = \{\omega_i : i = 1,2, \ldots, c\}$ be the finite set of $c$ classes. Let $\mathcal{X}_i \in \omega_i$ be the $i$th class set having $n_i$ number of training samples and $\mathcal{X}_1 \cup \mathcal{X}_2 \ldots \mathcal{X}_{c-1} \cup \mathcal{X}_c = \mathcal{X}$. If the set $\mathcal{X}$ is processed through a feature selection method $f(.)$ then it will give feature subset $\hat{\mathcal{X}} = f(\mathcal{X})$, where $\hat{\mathcal{X}}$ is in a $h$-dimensional feature space ($h < d$). To get the optimum value of $h$ let us consider a two-class case illustrated in Fig. 1. In the figure the two oval shapes denote the

training sets $\not{x}_1$ and $\not{x}_2$. A classifier is used to separate the feature space into two regions namely $\mathrm{R}_1$ and $\mathrm{R}_2$. The probability of samples correctly labeled is denoted by $P_{r1}$ and $P_{r2}$. The probability of samples given a class is denoted by $P_{x1}$ and $P_{x2}$. The error of misclassification is denoted by $\varepsilon$. The probabilities $P_{r1}$, $P_{r2}$, $P_{x1}$ and $P_{x2}$ can be given as

$$\mathrm{P}_{r1} = \int_{\mathrm{R}_1} p(\mathbf{x}|\omega_1)P(\omega_1)d\mathbf{x}, \quad \mathrm{P}_{r2} = \int_{\mathrm{R}_2} p(\mathbf{x}|\omega_2)P(\omega_2)d\mathbf{x}$$

and $$\mathrm{P}_{x1} = \int_{\not{x}_1} p(\mathbf{x}|\omega_1)P(\omega_1)d\mathbf{x}, \quad \mathrm{P}_{x2} = \int_{\not{x}_2} p(\mathbf{x}|\omega_2)P(\omega_2)d\mathbf{x}$$

where $p(\mathbf{x}|\omega_i)$ is the class-conditional probability density function and $P(\omega_i)$ is the a priori probability. The error $\varepsilon$ can be evaluated by

$$\varepsilon = P_{x1} + P_{x2} - (P_{r1} + P_{r2}).$$

It is obvious that error in different dimensional feature space would be different. Let the error is represented in $h$-dimensional feature space and let extending it for a $c$ class case, we get

$$\varepsilon_h = \sum_{i=1}^{c} \int_{\not{x}_i} p(\hat{\mathbf{x}}|\omega_i)P(\omega_i)d\hat{\mathbf{x}} - \sum_{i=1}^{c} \int_{\mathrm{R}_i} p(\hat{\mathbf{x}}|\omega_i)P(\omega_i)d\hat{\mathbf{x}} \qquad (1)$$

where $\hat{x} \in \hat{\not{x}}$. If the features are ranked using the feature selection method $f(.)$ then top $h$ features can be used for which $\varepsilon_h$ is minimum. For gene expression profile we can approximate equation 1 as

$$\varepsilon_h = \sum_{i=1}^{c} \sum_{\not{x}_i} p(\hat{\mathbf{x}}|\omega_i)P(\omega_i) - \sum_{i=1}^{c} \sum_{\mathrm{R}_i} p(\hat{\mathbf{x}}|\omega_i)P(\omega_i) \qquad (2)$$

When $\varepsilon_h = 0$ at $h$, there will be no overlapping between samples of different classes. In situations where the computation of class-conditional probability density function is extremely tedious or not possible, a simpler error function could be applied,

$$\varepsilon'_h = n - \sum_{i=1}^{c} number\ of\ samples\ belongs\ to\ R_i\ given\ \omega_i. \qquad (3)$$

*Results:* Three DNA microarray gene expression datasets are used. The datasets are described in Table 1. We have used nearest centroid classifier (NCC) to find the regions $\mathrm{R}_i$. The proposed strategy has been applied on two feature ranking methods namely information gain (InfoGain) and SVM to rank the genes. The choosing of value $h$ is illustrated in Figure 2 on SRBCT dataset. Here, the range for minimum and stable error is between 37 and 63. Therefore, we selected $h = 37$. The classification accuracy of several methods has been compared in Table 2, 3 and 4 for SRBCT dataset, MLL Leukemia dataset and Lung Cancer dataset respectively. In all datasets, the proposed strategy is able to achieve a test error rate at least equivalent if not better than current state of the art methods. It is noteworthy that in one case the proposed strategy achieves this good performance with up to 500 times less features than other method. Having a smaller subset of genes would give biologists a better chance in finding and/or understanding pathways that are important in the disease.

*Conclusion:* In this paper, we have presented a strategy for finding the minimum number of genes from gene expression dataset to achieve the high classification accuracy. This strategy has strong theoretical basis and displays promising results empirically.

**Reference**

1 Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., Gene selection for cancer classification using support vector machines, Machine Learning., 46 389–422, 2002

2 Yu, L. and Liu, H., Efficient feature selection via analysis of relevance and redundancy. J. Machine Learning Research, 5, 1205–1224, 2004.

3 Jafari, P. and Azuaje, F., An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. BMC Med. Inform. Decision Making, 6, 27, 2006.

4 Mamitsuka, H., Selecting features in microarray classification using ROC curves. Pattern Recognition, 39, pp. 2393–2404, 2006.

5 Tao L., Zhang, C. and Ogihara, M., “A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression”, Bioinformatics, vol, 20, no. 14, pp. 2429-2437, 2004

6 Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C. and Meltzer, P.S. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural network. Nature Medicine, vol. 7, pp. 673-679, 2001

7 Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., and Korsemeyer, S.J., MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genetics, vol. 30, pp 41-47, 2002.

8 Gordon, G.J., Jensen, R.V., Hsiao, L.-L., Gullans, S.R., Blumenstock, J.E., Ramaswamy, S., Richards, W.G., Sugarbaker, D.J., Bueno, R.: Translation of Microarray Data Into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma, Cancer Research, vol. 62, pp 4963-4967, 2002.

9 Tan A.C., Gilbert D., Ensemble machine learning on gene expression data for cancer classification, *Appl. Bioinformatics*, 2(3 Suppl), pp. S75-83, 2003.

10 Li J., Wong L., Using rules to analyse bio-medical data: a comparison between C4.5 and PCL, In: *Advances in Web-Age Information Management.* Berlin / Heidelberg: Springer, pp. 254-265, 2003.

11 Cong G., Tan K.-L., Tung A.K.H., Xu X., Mining top-k covering rule groups for gene expression data. In: *the ACM SIGMOD International Conference on Management of Data,* pp. 670-681, 2005.

**Authors’ affiliations:** Alok Sharma, Chuan H. Koh, Seiya Imoto and Satoru Miyano (Laboratory of DNA Information Analysis, Human Genome Center, University of Tokyo, Japan). Alok Sharma also with the University of the South Pacific, Fiji.

**Figure captions:**

Figure 1 An illustration using two-class case.

Figure 2 Selection process for the optimum value $h$.

**Table captions:**

Table 1 DNA microarray gene expression datasets

Table 2 Comparison of strategies on SRBCT dataset

Table 3 Comparison of strategies on MLL Leukemia dataset

Table 4 Comparison of strategies on Lung Cancer dataset

Figure 1

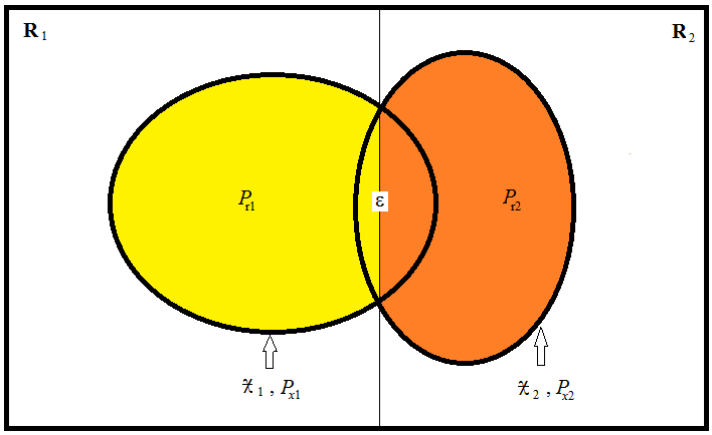


Figure 2

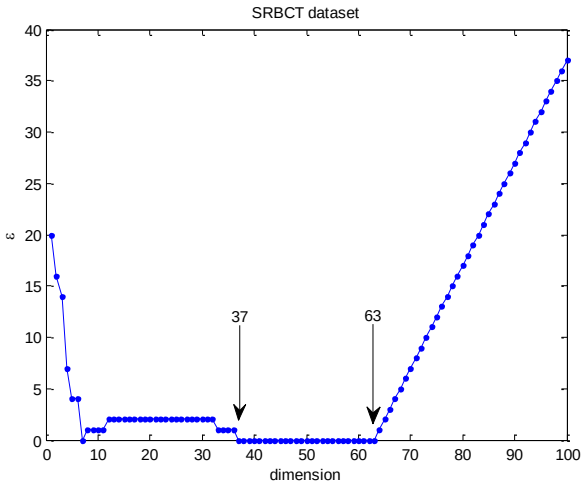

Table 1

| **Datasets** | **Class** | **Number of features** | **Number of training samples** | **Number of testing samples** |
|---|---|---|---|---|
| SRBCT [6] | 4 | 2308 | 63 | 20 |
| MLL Leukemia [7] | 3 | 12582 | 57 | 15 |
| Lung Cancer [8] | 2 | 12533 | 32 | 149 |

Table 2

| **Methods (Feature Selection + Classification)** | **# selected genes** | **SRBCT** (Classification accuracy on test data) |
|---|---|---|
| InfoGain + SVM 1 vs all [5] | 150 | 95% |
| One-dimensional SVM + SVM Naïve Bayes [5] | 150 | 63% |
| One-dimensional SVM + SVM random [5] | 150 | 91% |
| One-dimensional SVM + SVM exhaustive [5] | 150 | 95% |
| Proposed strategy + InfoGain + nearest centroid classifier | 37 | 100% |
| Proposed strategy + InfoGain + nearest neighbor classifier | 37 | 100% |
| Proposed strategy + SVM + nearest centroid classifier | 10 | 90% |
| Proposed strategy + SVM + nearest neighbor classifier | 10 | 90% |

Table 3

| **Methods (Feature Selection + Classification)** | **# selected genes** | **MLL Leukemia** (Classification accuracy on test data) |
|---|---|---|
| SVM + SVM random [5] | 150 | 100% |
| InfoGain + Naïve Bayes [5] | 150 | 54% |
| One-dimensional SVM + SVM random [5] | 150 | 100% |
| One-dimensional SVM + SVM exhaustive [5] | 150 | 100% |
| Proposed strategy + InfoGain + nearest centroid classifier | 46 | 93.3% |
| Proposed strategy + InfoGain + nearest neighbor classifier | 46 | 86.7% |
| Proposed strategy + SVM + nearest centroid classifier | 37 | 93.3% |
| Proposed strategy + SVM + nearest neighbor classifier | 37 | 100% |

Table 4

| **Methods (Feature Selection + Classification)** | **# selected genes** | **Lung Cancer** (Classification accuracy on test data) |
|---|---|---|
| Discretization + decision trees [9] | 5365 | 93% |
| Boosting [10] | unknown | 81% |
| Bagging [10] | unknown | 88% |
| RCBT [11] | 10-40 | 98% |
| Proposed strategy + InfoGain + nearest centroid classifier | 10 | 99.3% |
| Proposed strategy + InfoGain + nearest neighbor classifier | 10 | 99.3% |
| Proposed strategy + SVM + nearest centroid classifier | 12 | 98.7% |
| Proposed strategy + SVM + nearest neighbor classifier | 12 | 100% |