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

A. Sharma, C.H. Koh, S. Imoto and S. 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 the dimensionality problem and improves the interpretability of the problem. Numerous feature selection methods have been proposed in the 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 arbitraryly or heuristic fashion. Proposed is a theoretical way to determine the optimal number of genes to be selected for a given task. This proposed strategy has been applied on a number of gene expression datasets and promising results have been obtained.

*Introduction:* Dimensionality reduction techniques are applied to high dimensional problems for reducing computational complexity and improving generalisation 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 a linear or a nonlinear 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 characterising heterogeneity of human cancers.

Numerous feature selection methods have been proposed in the literature [1–4]. A comprehensive study can be found in [5]. These methods explore the significance of genes and rank them based on a certain feature score. Then, the top $h$ genes are selected for downstream applications 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 that the chosen $h$ is too large and a much lower $h$ would achieve similar or even better performance. In this Letter, 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.

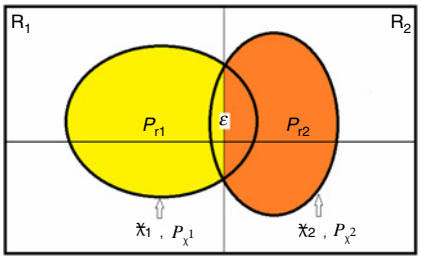


**Fig. 1** *Illustration using two-class case*

*Proposed strategy:* The mathematical notations used in this Letter are defined as follows. Let $\chi = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ 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 $\chi_i \in \omega_n$ be the $i$th class set having $n_i$ number of training samples and $\chi_1 \cup \chi_2 \ldots \chi_{c-1} \cup \chi_c = \chi$. If the set $\chi$ is processed through a feature selection method $f(.)$ then it will give feature subset $\hat{\chi} = f(\chi)$, where $\hat{\chi}$ is in an $h$-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 $\chi_1$ and $\chi_2$. A classifier is used to separate the feature space into two regions namely $R_1$ and $R_2$. The probability of samples correctly labelled 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

$$P_{r1} = \int_{R_1} p(\mathbf{x}|\omega_1)P(\omega_1)d\mathbf{x}, \quad P_{r2} = \int_{R_2} p(\mathbf{x}|\omega_2)P(\omega_2)d\mathbf{x}$$

and

$$P_{x1} = \int_{\chi_1} p(\mathbf{x}|\omega_1)P(\omega_1)d\mathbf{x}, \quad P_{x2} = \int_{\chi_2} p(\mathbf{x}|\omega_2)P(\omega_2)d\mathbf{x}$$

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

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

It is obvious that error in different dimensional feature space would be different.

Let the error be represented in $h$-dimensional feature space, and extending it for a $c$ class case, we get

$$\varepsilon_h = \Sigma_{i=1}^{c} \int_{\chi_i} p(\hat{\mathbf{x}}|\omega_i)P(\omega_i)d\hat{\mathbf{x}} - \Sigma_{i=1}^{c} \int_{R_i} p(\hat{\mathbf{x}}|\omega_i)P(\omega_i)d\hat{\mathbf{x}} \quad (1)$$

where $\hat{x} \in \hat{\chi}$. If the features are ranked using the feature selection method $f(.)$ then the top $h$ features can be used for which $\varepsilon_h$ is minimum. For gene expression profile we can approximate (1) as

$$\varepsilon_h = \Sigma_{i=1}^{c} \Sigma_{\chi_i} p(\hat{\mathbf{X}}|\omega_i)P(\omega_i) - \Sigma_{i=1}^{c} \Sigma_{\chi_i} p(\hat{\mathbf{X}}|\omega_i)P(\omega_i) \quad (2)$$

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:

$$\acute{\varepsilon}_h = n - \Sigma_{i=1}^{c} \text{number of samples belongs to } R_i \text{ given } \omega_i \quad (3)$$

**Table 1:** DNA microarray gene expression datasets

| 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 |

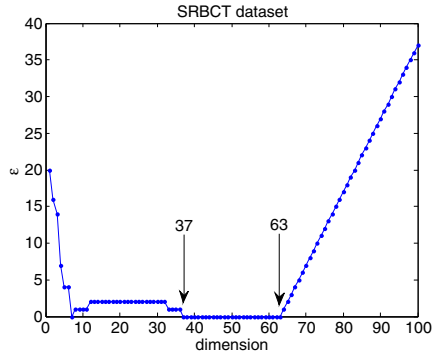


**Fig. 2** *Selection process for optimum value of h*

**Table 2:** Comparison of strategies on SRBCT dataset

| Methods (feature selection + classification) | Number of selected genes | SRBCT (classification accuracy on test data) (%) |
|---|---|---|
| InfoGain + SVM 1 versus 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 neighbour classifier | 37 | 100 |
| Proposed strategy + SVM + nearest centroid classifier | 10 | 90 |
| Proposed strategy + SVM + nearest neighbour classifier | 10 | 90 |

*Results:* Three DNA microarray gene expression datasets are used. The datasets are described in Table 1. We have used the nearest centroid classifier (NCC) to find the regions $R_i$. 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 Fig. 2 on the 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 Tables 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 to, 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 methods. Having a smaller subset of genes would give biologists a better chance of finding and/or understanding pathways that are important in the disease.

**Table 3:** Comparison of strategies on MLL Leukemia dataset

| Methods (feature selection + classification) | Number of 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 neighbour classifier | 46 | 86.7 |
| Proposed strategy + SVM + nearest centroid classifier | 37 | 93.3 |
| Proposed strategy + SVM + nearest neighbour classifier | 37 | 100 |

**Table 4:** Comparison of strategies on Lung Cancer dataset

| Methods (feature selection + classification) | Number of selected genes | Lung cancer (classification accuracy on test data) (%) |
|---|---|---|
| Discretisation + 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 neighbour classifier | 10 | 99.3 |
| Proposed strategy + SVM + nearest centroid classifier | 12 | 98.7 |
| Proposed strategy + SVM + nearest neighbour classifier | 12 | 100 |

*Conclusion:* We present a strategy for finding the minimum number of genes from a gene expression dataset to achieve high classification accuracy. This strategy has a strong theoretical basis and displays promising results empirically.

*Acknowledgment:* This work was partly supported by Grant-in-Aid for JSPS Fellows (22-00364).


*24 February 2011*
doi: 10.1049/el.2011.0526
One or more of the Figures in this Letter are available in colour online.

A. Sharma, C.H. Koh, S. Imoto and S. Miyano (*Laboratory of DNA Information Analysis, Human Genome Center, University of Tokyo, Japan*)

## References

1 Guyon, I., Weston, J., Barnhill, S., and Vapnik, V.: 'Gene selection for cancer classification using support vector machines', *Mach. Learn.*, 2002, **46**, pp. 389–422
2 Yu, L., and Liu, H.: 'Efficient feature selection via analysis of relevance and redundancy', *J. Mach. Learn. Res.*, 2004, **5**, pp. 1205–1224
3 Jafari, P., and Azuaje, F.: 'An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors', *BMC Med. Inf. Decision Making*, 2006, **6**, (art. id. 27)
4 Mamitsuka, H.: 'Selecting features in microarray classification using ROC curves', *Pattern Recognit.*, 2006, **39**, pp. 2393–2404
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*, 2004, **20**, (14), pp. 2429–2437
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 Med.*, 2001, **7**, pp. 673–679
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 Genet.*, 2002, **30**, pp. 41–47
8 Gordon, G.J., Jensen, R.V., Hsiao, L.-L., Gullans, S.R., Blumenstock, J.E., Ramaswamy, S., Richards, W.G., Sugarbaker, D.J., and Bueno, R.: 'Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma', *Cancer Res.*, 2002, **62**, pp. 4963–4967
9 Tan, A.C., and Gilbert, D.: 'Ensemble machine learning on gene expression data for cancer classification', *Appl. Bioinformat.*, 2003, **2**, (3 Supplement), pp. S75–83
10 Li, J., and Wong, L.: 'Using rules to analyse bio-medical data: a comparison between C4.5 and PCL' *in* 'Advances in web-age information management' (Springer, Berlin/Heidelberg, Germany, 2003), pp. 254–265
11 Cong, G., Tan, K.-L., Tung, A.K.H., and Xu, X.: 'Mining top-k covering rule groups for gene expression data'. ACM SIGMOD Int. Conf. on Management of Data, Baltimore, MD, USA, 2005, pp. 670–681