行政院國家科學委員會專題研究計畫 成果報告

利用本質相關係數建立樣本分類法則之可行性評估
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 95-2119-M-002-045-
執 行 期 間 ：95 年 08 月 01 日至 96 年 07 月 31 日
執 行 單 位 ：國立臺灣大學農藝學系暨研究所

計 畫 主 持 人：劉力瑜

計畫參與人員：碩士班研究生-兼任助理：蕭雅純、馬梓豪

處 理 方 式 ：本計畫可公開查詢

中 華 民 國 96 年 10 月 31 日

行政院國家科學委員會補助專題研究計畫成果報告

利用本質相關係數建立樣本分類法則之可行性評估

計畫類別：個別型計畫
計畫編號：NSC 95－2119－M－002－045－
執行期間：2006 年 08 月 01 日至 2007 年 07 月 31 日

計畫主持人：劉力瑜
計畫參與人員：碩士班研究生-兼任助理：蕭雅純、馬梓豪

成果報告類型：精簡報告

處理方式：本計畫可公開查詢

執行單位：國立台灣大學農藝學系暨研究所

中　華　民　國　96 年　10 月　27　日

中文摘要

依據各試驗單位的特性建立樣本分類法則, 可細分爲兩個步驟: (1) 篩選數個可明確判別群集的變數, (2) 利用選擇的變數建立最佳分類法則。其中步驟 (1) 通稱爲特徵篩選 (feature selection), 當可供做爲分類依據的變數個數很多時, 選擇少數幾個足以適當分類樣本的變數除了有助於降低成本, 挑選適當的變數也是準確分類樣本的關鍵 (Sima et al., 2005)。變數篩選標準可歸類爲相關性度量與錯分率度量兩大類。錯分率度量與分類模式有關, 不同分類模式結果相異。相關性度量則取決於關聯性統計值的選擇。Hsing 等人於 2005 的論文中提出本質相關係數 (coefficient of intrinsic dependence, 簡稱 CID), 藉由量化變數邊際分布與條件分布間的差異來描述變數間的相關程度, 屬於關聯性統計值的一員, 且不針對描述特定相關模式, 適合做爲特徵篩選的依據。Hsing 等人於 2005 的論文也提到, 在小樣本時, CID 估值仍舊能忠實反應變數間實際相關程度。本研究目的即在探討利用 CID 建立小樣本分類法則的可行性, 並與傳統分類法互相比較其優劣。最後, 此一方法將實際應用在乳癌基因表現資料的分類。

關鍵詞: 本質相關係數, 分類法則, 特徵篩選, 微陣列

Abstract

The problem of classification is to assign objects to one of the mutually exclusive subgroups in the population based on the object's characteristics. To build a precise rule of classification, a two-step procedure is usually performed on the training dataset: (1) selecting a few features that are most informative in the sense of decision making; (2) deriving the formula that outputs optimal allocation of objects. Selecting appropriate features is particularly essential for a successful classification. Recent methods of feature selection consider either the misclassification rate of objects given information of a set of variables, or the association between variables and class label. The former yields inconsistent results for different settings of classifiers. The later is subject to the choice of association measures. The analysis of actual data from a study of breast cancer gene expression is included.

Hsing et al. (2005) has proposed a new measure of association, the coefficient of intrinsic dependence, or CID. The CID captures not only linear but general association among variables. It was also demonstrated that CID is capable of putting variables in appropriate order according to their degree of association to the target variable even when sample size is small. This research will broaden the work of Hsing et al. (2005) by applying CID in feature selection. It will be followed by construction of Bayes classifiers and comparisons to conventional methods.

Keywords: CID, classification, feature selection, microarray

1 Introduction

The problem of classification is to assign objects to one of the mutually exclusive subgroups in the population based on the object's characteristics. The methodology of classification has a long history of development (Jain et al., 2000). To build a classification rule, a two-step procedure is usually performed on the training dataset: (1) selecting a few features that distinguish classes the most; (2) deriving the decision-making formula, or the classifier, that best allocates objects. The task in the first step is usually referred as feature selection. Given a set of variables, feature selection aims to select a variable subset that performs the best as to certain statistical or mechanical criteria. This procedure not only reduces cost of trials but increases the accuracy of classification (Jain and Zongker, 1997). It

was also noted feature selection is particularly essential for proper classification (Sima et al., 2005).

Conventional methods of feature selection fall into two categories. One category recognizes variables that best determine the class labels in the training dataset and collects the subset of variables making the least error of allocation. It is intuitive to consider misclassification rate, but it can be accessed only after the classifier is built. To proceed feature selection, one must decide a specific type of classifier. An exhaustive search is then conducted toward all possible combinations of k features and estimation of misclassification rate follows. The number of variables, k, is usually pre-determined as well. Sima et al. (2005) illustrated the choice of classifiers has little effect on classification results. It is the algorithms of error estimation that matters. The precision of error estimation algorithms rapidly corrupt as sample size decreases. As a result, an improper classifier may be wrongly concluded.

The second category of feature selection discerns variables whose values alter from one class to another. In the other words, the variables that are associated with the class labels are favored. Various statistical methods can be performed to determine whether the values of a particular feature differentially expressed. For example, Student's t-test is widely used when there are only two classes of interest. Welch's approximation is alternatively suggested when assuming the population variances of the two classes are unequal. One would prefer a nonparametric method, such as Wilcoxon rank-sum test, if he is concerned about the validation of normality assumption for Student's t-test.

Hsing et al (2005) has proposed a new measure of association - the coefficient of intrinsic dependence, or CID. The CID takes any real values between 0 and +1 inclusive. It is +1 in the case of full dependence and is zero in the case of independence. As the level of dependence ascends, the CID value goes from 0 to 1. Naturally, CID is applicable in feature selection because it follows the convention of association measures. Hsing et al. (2005) has demonstrated merits of CID in feature selection especially under small sample. Their results are confirmed by simulation studies. One of the objectives of this research is to continue the work of Hsing et al. (2005) on development of classifier. We will compare the feature selection results and the misclassification rates to that of conventional methods. We will exercise CID on a complete breast cancer gene expression data that were produced by National Taiwan University Hospital.

2 Method

2.1 Coefficient of Intrinsic Dependence

Let W and Z be explanatory and target variables, respectively. The CID of Z given W is defined as follow:

$$\mathrm{CID}(Z|W) = \frac{\int_0^1 \mathrm{Var}[\mathrm{E}(I(G(Z) \leq v)|W)]dv}{\int_0^1 \mathrm{Var}[I(G(Z) \leq u)]du}, \tag{1}$$

where $G(\cdot)$ is the marginal cdf of Z, and $I(A)$ is an indicator function such that

$$I(A) = \begin{cases} 1 & \text{if } A \text{ is true;} \\ 0 & \text{otherwise.} \end{cases}$$

It was shown that the numerator integrates the squared distance between the marginal cdf of Z and the conditional cdf of Z given W. By variance decomposition, the denominator serves to standardize the value of CID in [0,1]. When W and Z are nearly independent, the knowledge of W provides little information about Z. The conditional and marginal distributions of Z are therefore alike to each other, which makes the numerator of CID nearly 0. In the other hand, if two variables are highly relevant, one can easily discriminate the object only by the knowledge of explanatory variables. In these cases, CID yields values close to 1. In summary, CID has following properties:

1. CID always has a value between 0 and 1. If two random variables W and Z are fully dependent, $\text{CID}(Z|W) = \text{CID}(W|Z) = 1$. In the other hand, $\text{CID}(Z|W) = \text{CID}(W|Z) = 0$ if W and Z are independent to each other.

2. The causal relationship between variables is taken into account by asymmetric property of CID. That is, $\text{CID}(Z|W)$ might be different from $\text{CID}(W|Z)$.

3. CID requires no distributional assumptions and is invariant under transformations of variables.

4. It is ready to implement in different occasions, such as numerical, categorical, or multivariate cases, by inserting appropriate distribution functions.

In the case of classification, the target variable is the class of objects, which is denoted as Y. When allocating objects into two classes, which is assumed in this project, there are only two possible results of observed values of Y: 0 or 1. Then a simpler version of (1) can be derived:

$$\begin{aligned} \text{CID}(Y|\mathbf{x}) &= \frac{\text{Var}[\text{E}(Y|\mathbf{x})]}{\text{Var}(Y)} \\ &= 1 - \frac{\text{E}[\text{Var}(Y|\mathbf{x})]}{\text{Var}(Y)} \end{aligned} \tag{2}$$

In particular, if $\mathbf{x}$ is univariate with binary outcomes then CID is equal to the square of correlation coefficient of Y and $\mathbf{x}$.

In the absence of knowledge about true distribution functions, the denominator and numerator of CID are separately estimated from a sample of size n. If $\mathbf{x}$ involves continuous random variable(s), a binning process can be employed to estimate the conditional distributions. We firstly determine the number of subspaces, B, according to experience. Then the cutting points are decided in a way that each subspace contains approximately equal number of observations. The estimate of CID is

$$\widehat{\text{CID}}(Y|\mathbf{x}) = 1 - \frac{\sum_{i=1}^{B} \hat{p}_i(1-\hat{p}_i) \times n_i/n}{\hat{\pi}(1-\hat{\pi})} \tag{3}$$

where

$$\begin{aligned} \hat{\pi} &= \sum_{j=1}^{n} Y_j/n; \\ n_i &= \sum_{j=1}^{n} I(\mathbf{x}_j \in A_i); \\ \hat{p}_i &= \sum_{j=1}^{n} Y_j \times I(\mathbf{x}_j \in A_i)/n_i; \\ A_i &= \text{the } i\text{th defined subspace} \end{aligned}$$

2.2 Conventional Statistical Methods

We compare the feature selection results of CID with that of the conventional statistics described in this section. Let $x_1, \cdots, x_{n_1}$ be a random sample taken from the first distribution with mean μ_1 and variance σ_1^2 and $y_1, \cdots, y_{n_2}$ be a random sample taken from the second distribution with mean μ_2 and variance σ_2^2. Their sample means and variances are denoted as $\bar{x}$, $\bar{y}$, S_x^2, S_y^2, repsectively.

2.2.1 Student's *T*-Statistics and Welch's Approximation

If assuming $\sigma_1^2 = \sigma_2^2$, the statistic can be written as

$$t_0 = \frac{\bar{x} - \bar{y}}{\sqrt{S_p^2(1/n_1 + 1/n_2)}},$$

where S_p^2 is the pooled estimate of the common variance. If $\sigma_1^2 \neq \sigma_2^2$, we adopt the Welch Approximation

$$t_0 = \frac{\bar{x} - \bar{y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}.$$

The features are ranked based on the absolute values of t_0. Features having the largest absolute values of t_0 are claimed to be the most differentially expressed.

2.2.2 Wilcoxon Rank-Sum Test

We combine two samples $x_1, \cdots, x_{n_1}$ and $y_1, \cdots, y_{n_2}$ and rank the observations in the combined sample from the smallest (1) to the largest $(n_1 + n_2)$. The ranks of $x_1, \cdots, x_{n_1}$ are denoted as $r_{x_1}, \cdots, r_{x_{n_1}}$ and the ranks of $y_1, \cdots, y_{n_2}$ are denoted as $r_{y_1}, \cdots, r_{y_{n_2}}$. Let

$$T_1 = \sum_i r_{x_i}, \quad T_2 = \sum_i r_{y_i},$$

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_1, \text{ and}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_2.$$

For a two-tailed test, the test statistic $U = \min\{U_1, U_2\}$; for a one-tailed test, the test statistic $U = U_1$. The features are ranked based on the values of U. The feature associated with higher value of U is more differentially expressed.

2.3 Rank Statistics

The CID and conventional methods respectively rank features from simulated samples. We compare ranking list produced by each method to the true ranking. Those methods retrieve the most original ranking are determined to be the best. It is intuitive to compute the correlation between the true ranking and the ranking list yielded from the sample. If the method well retrieve the original ranking then the correlation should be high. There is one thing we want to draw readers' attention to. In a massive dataset such as microarray expression data, people aims to find only a few essential features for classification to achieve dimensional reduction. Unimportant features may show up and down simply because of random error. Therefore, we do not hesitate to omit ranking unimportant features. Only the subset of features proclaimed important are of interest. Braga-Neto et al. (2004) suggested two ranking statistics for comparison of ranking subsets. They would be adopted in our study to evaluate the performance of each feature-selection criterion (i.e., CID, t/Welch test, or Wilcoxon rank sum test). They are described as below. Let k be the true ranking (most important ones on the top) and k^* be the estimated one.

- The first statistic calculates the mean absolute difference between the top K features in the true list and the top K features in the estimated list:

$$T_1 = \frac{\sum_{i=1}^{K} |k - k^*|}{K}.$$

- The second statistics counts how many top K features in the true list are also among the top K features in the estimated list:

$$T_2 = \sum_{i=1}^{K} I(k^* \leq K),$$

where I is the indicator function.

2.4 Bayes Decision Theory

For known class conditional densities $p_k(\mathbf{x}) = \Pr(\mathbf{x}|Y = k)$ and class priors π_k, let

$$p(k|\mathbf{x}) = \frac{\pi_k p_k(\mathbf{x})}{\pi_1 p_1(\mathbf{x}) + \pi_2 p_2(\mathbf{x})}$$

denote the posterior probability of class $k(k = 0, 1)$ given $\mathbf{x}$. The Bayes rule, d_B, classify an object with observed feature $\mathbf{x}$ as that j for which the posterior probability is maximum:

$$d_B(\mathbf{x}) = j = \arg\max_k p(k|\mathbf{x}).$$

In other words, the object is allocated to the class that is more possible to occur given the values of $\mathbf{x}$. The Bayes misclassification rate, denoted as R_B can be estimated as follow:

$$\begin{aligned} &\Pr(d_B(\mathbf{x}) \neq Y) \\ &= \pi_0 \Pr(d_B(\mathbf{x}) = 1|Y = 0) \\ &+ \pi_1 \Pr(d_B(\mathbf{x}) = 0|Y = 1). \end{aligned} \tag{4}$$

Bayes rule has one nice property that, when relevant distributions and priors are known, it produces the least errors than any other classifiers.

One wishes to allocate the object based on observed values of features that are corresponding to high CID values, $\mathbf{x}^*$. Suppose $\mathbf{x}^* \in A_i$, the probability that the object belonging to either group has already estimated while computing the CID value on training data. Hence the Bayes rule can be immediately implanted for classification. The objects is assigned to group 1 if $p_i > 0.5$ and to group 0 otherwise.

2.5 Data Simulation

We will compare feature selection results of CID to that of the other conventional methods. All methods will be tested on a simulated data whose true Bayes errors can be accessed. Three models are used to generate samples: the Gaussian model with unequal means and variances, and the log-normal model. They are described as follow.

2.5.1 The Gaussian Model with Unequal Means

Suppose all samples come from one of two equally likely p-dimensional Gaussian distributions: $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$; both μ_1 and μ_2 are $p \times 1$ vectors and Σ is a $p \times p$ positive-definite matrix. To simplify the scenario, let $\Sigma =$ identity matrix, $\mu_1 = -\delta\mathbf{a}$, and $\mu_2 = \delta\mathbf{a}$. The true error rate $R_B(\mathbf{x})$ decreases as the value of δ increases.

2.5.2 The Gaussian Model with Unequal Variances

Suppose a particular feature performs similar in average in two classes but its values are more varied in one of the classes. In the classification point of view, this feature can be used to discriminate objects especially when the variances are highly unequal since the two distributions do not much overlap. Let all samples come from one of two equally likely p-dimensional Gaussian distributions: $N(\mu, \Sigma_1)$ and $N(\mu, \Sigma_2)$, where μ is a $p \times 1$ vector and Σ_1 and Σ_2 are $p \times p$ positive-definite matrices. To simplify the scenario, let $\mu_1 = \mathbf{0}$, $\Sigma_1 =$ identity matrix, and $\Sigma_2 = diag(\sigma_1^2, \cdots, \sigma_p^2)$, where $\sigma_i^2 = 1 + \frac{(i-1)(\delta-1)}{p-1}$. The true error rate $R_B(\mathbf{x})$ decreases as the value of δ increases.

2.5.3 The Lognormal Model

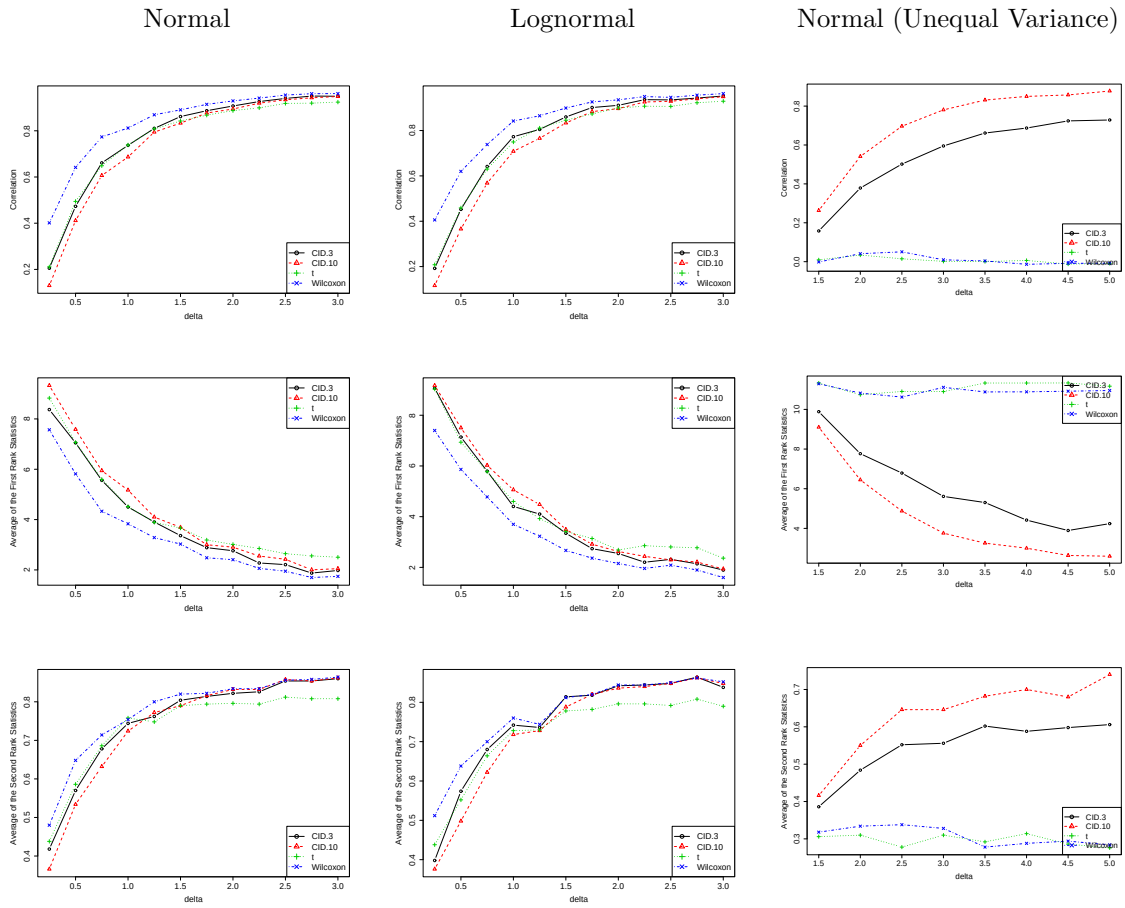
All samples are drawn from the Gaussian model with unequal means as described in Section 2.5.1 and take exponentiation. Therefore, the true error rate $R_B(\mathbf{x})$ also decreases as the value of δ increases.

2.6 Clinical Breast Cancer Expression Array

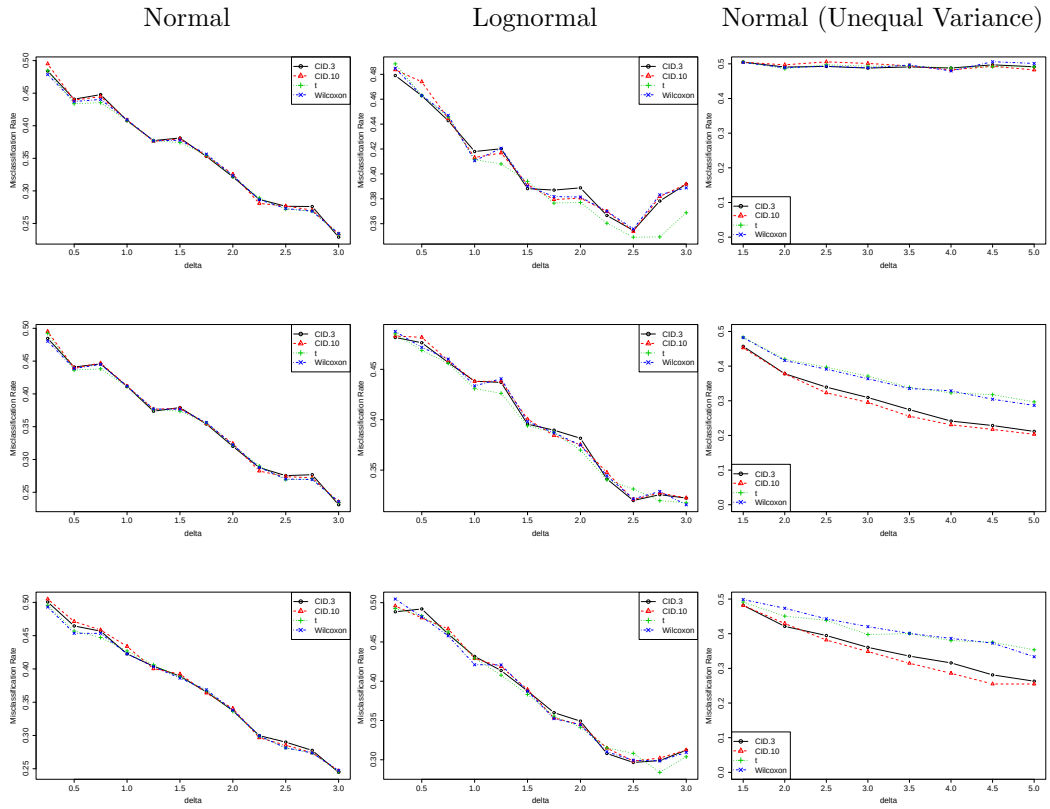
Total 80 clinical arrays were from a patient cohort collected at National Taiwan University Hospital (NTUH). These clinical arrays were generated using the Human 1A (version 2) oligonucleotide microarray (Agilent technologies) according to the methods provided by the manufacture (Lien HC et al. 2007). All patients had given informed consent according to guidelines approved by the Institutional Review Board (IRB) of NTUH. Paraffin section of breast cancer specimens were stained with CONFIRM anti-Estrogen Receptor (SP1) antibody (Ventana Medical System, Inc) using Ventana Autostainer (BenchMark LT Full System)(Ventana Medical Systems, Inc). The control slides for tumor specimens were stained using iVIEW DAB Detection kit (Ventana Medical System, Inc). All the ER immunochemical stained slides were further examined by two experienced pathologists. There are 18 ER(−), 60 six ER(+) and two unidentified specimens in the 80 clinical arrays.

3 Result

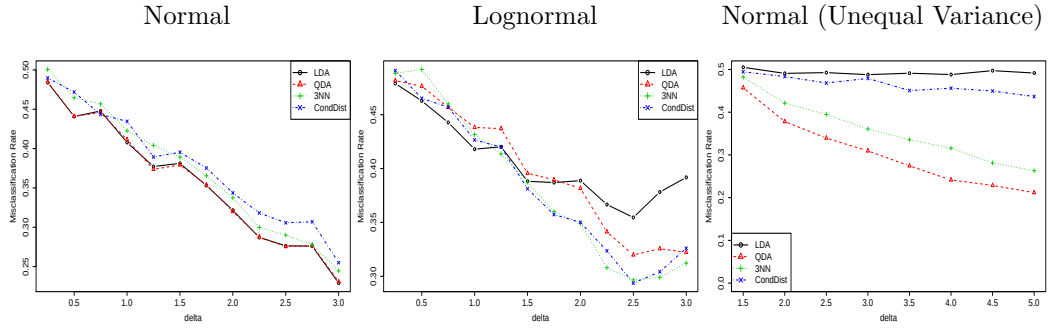
1. In the normal and lognormal model settings, t/Welch and Wilcoxon performed only slightly better than CID in the view of better retrieving the true ranking. However, t/Welch and Wilcoxon were not capable to identify features with unequal variances.



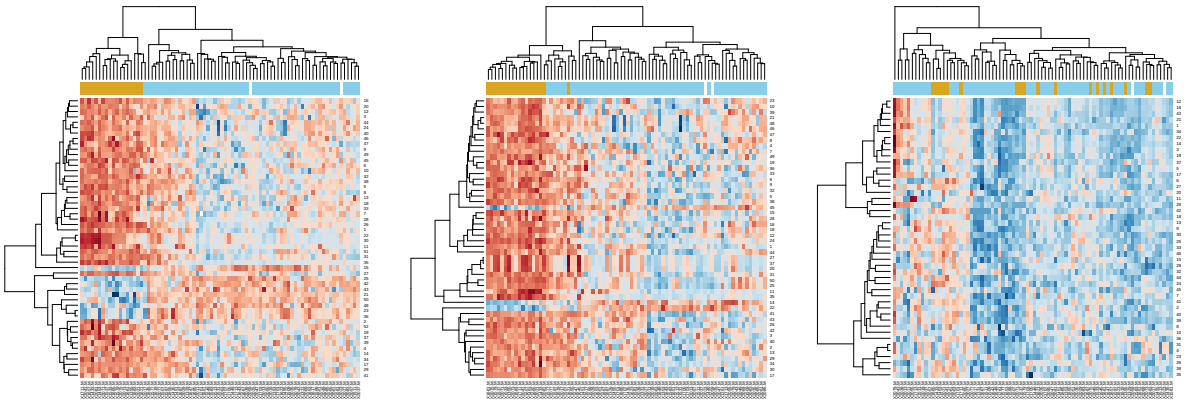
2. We use the top 3 features for each method to build the classifier by linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and 3-nearnest-neighbor method (3NN), respectively. The Figure below presented the misclassification rates for each method.



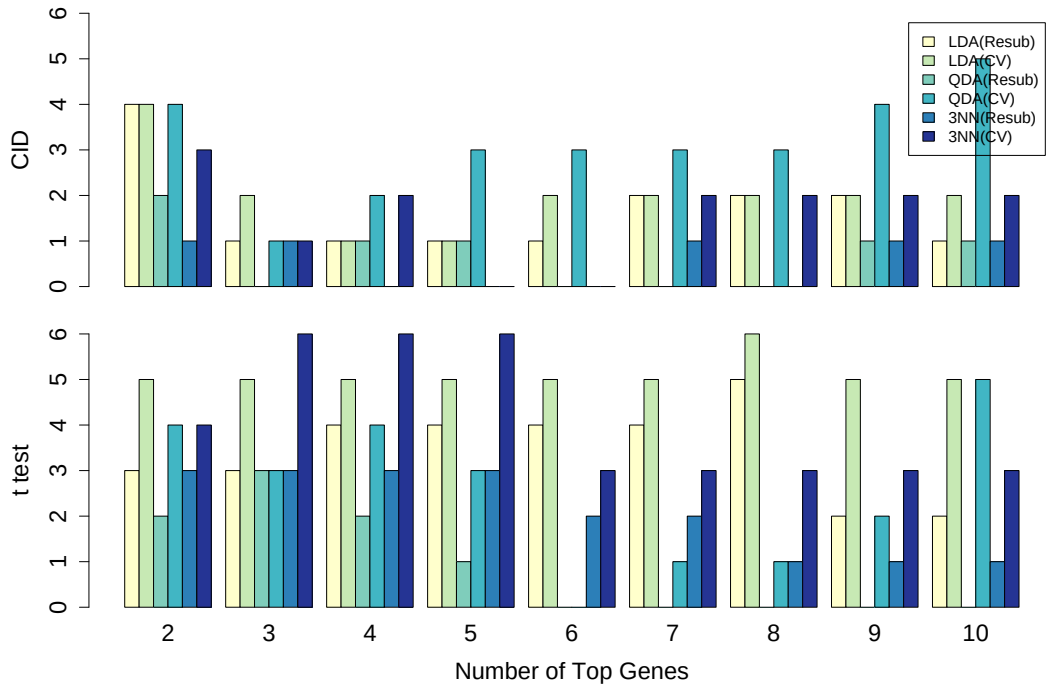
3. The probability that the object belonging to either group had estimated while computing the CID value on training data. The Bayes rule suggested to assign the object to group 1 if $p_i > 0.5$ and to group 0 otherwise. We use the top 3 features selected by CID to build the Bayes classifier. The classification results (CondDist) were compared with those of linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and 3-nearnest-neighbor method (3NN). The Bayes classifier performed remarkably better in the setting of lognormal model. As expected, LDA yielded the least misclassification rate for normal distributions with equal variance and QDA yielded the least misclassification rate for normal distributions with unequal variance. Another drawback of CID Bayes classifier was observed during the simulations. When the size of training data is small, we might not able to observe all possible combinations of categories of three top features. That caused problems to allocate the newly observed sample to one of the subset in the space as well as to obtain the estimated probability that the object belonging to either group.



4. Three feature selection methods were applied to 80 breast cancer clinical array data. Based on the top 50 features claimed by each method, we clustered the 80 specimens. The clustering results had been shown in the Figure below (from left to right: CID, t/Welch, and Wilcoxon). Observe that CID best separated ER+/− specimens (marked as brown for ER− and as blue for ER+ specimens below the top dendrogram). We also can speculate the two specimens with unknown ER status may be ER+.



5. By resubstitution and cross-validation we estimate the misclassification rate based on each method. Top 10 features were used to build the classifier. The Wilcoxon rank-sum test claimed that the top 50 feature are equally important so that Wilcoxon rank-sum test was excluded for this analysis. The Figure below showed that the features selected by CID produced less misclassification rate than that selected by t/Welch test.



4 Discussion

The high throughput techniques, such as microarrays, aid to monitor expression of thousands of genes simultaneously. One of the objective of microarray studies is to compare gene expression levels in two or several predetermined classes. Differential expression of genes can appear in different forms, for example, different means and/or variances. Therefore, the inquiry of a statistical method to universally identify different patterns of gene expressions arised.

In this study, we adopted the coefficient of intrinsic dependence (CID) to identify putative signatures for classification. Our results showed that CID is promising in supervised learning. The simulations had shown that CID is robust in selecting features with different means or different variances in two classes. When applying to the breast cancer clinical array data, the genes selected by CID best classified ER+/− patients.

However, CID is not appropriate to be immediately applied to unsupervised learning. Although the misclassification rate of CID was as low as those of conventional methods in most of the cases, CID suffered the curse of dimensionality the most. The small sample size relative to the number of variables (genes) is of particular concern in the microarray studies. When the sample size of the training data is small, one might yield a classifier that perfectly classifies the training sample but performs badly in the other samples. While applying CID in classification, the curse of dimensionality strikes from another direction. One may not observe a particular set of data in the training set but the same data appears in the test set. In this scenorio, the probability that the object belonging to certain group is not estimable. Another way to estimate the conditional distribution, such as nonparametric smoothing, might solve the problem.

References

1. U. Braga-Neto, R. Hashimoto, E.R. Dougherty, D.V. Nguyen, and R.J. Carroll (2004). Is Cross-Validation Better than Resubstitution for Ranking Genes? *Bioinformatics*. 20(2), 253-258.

2. T. Hsing, L.-Y. D. Liu, M. Brun, and E.R. Dougherty (2005). The coefficient of intrinsic dependence (feature selection using el CID). *Pattern Recognition*. 38: 623-636.

3. HC Lien, YH Hsiao, YS Lin, YT Yao, HF Juan, WH Kuo, MC Hung, KJ Chang, and FJ Hsieh (2007). Molecular signatures of metaplastic carcinoma of the breast by large-scale transcriptional profiling: identification of genes potentially related to epithelial-mesenchymal transition. *Oncogene*. 1-13.

4. A.K. Jain and D. Zongker (1997). Feature Selection - Evaluation, Application, and Small Sample Performance, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 19(2): 153-158.

5. A.K. Jain, R.P.W. Duin, and J. Mao (2000). Statistical Pattern Recognition: A Review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 22(1): 4-37.

6. I. Shmulevich, E.R. Dougherty, S. Kim, and W. Zhang (2001). Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*. 18, 261-274.

7. C. Sima, S. Attoor, U. Braga-Neto, J. Lowey, E. Suh, and E.R. Dougherty (2005). Impact of Error Estimation on Feature-Selection Algorithms. *Pattern Recognition*. 38(12): 2472-2482.

8. V.G. Tusher, R. Tibshirani, and G. Chu (2001). Significance Analysis of Microarrays Applied to the Ionizing Radiation Response. *Proceedings of the National Academy of Sciences*, 98(9):5116-5121.

計畫成果自評

由上述報告, 可以發現我們依據原計畫的設計進行資料模擬與分析, 並將方法應用在實際資料上, 確實達到計畫的預定目標, 所得結果對未來研究方向的設計也有很大助益。未來我們將會針對本篇觀察到的結果, 對方法學再進一步修正。我們已經將成果撰寫成技術報告, 目前正在潤飾的階段, 完成後將投稿至學術期刊。