Estimating Optimal Feature Subsets Using Efficient Estimation of High-Dimensional Mutual Information

Tommy W. S. Chow, *Senior Member, IEEE,* and D. Huang

***Abstract*—A novel feature selection method using the concept of mutual information (MI) is proposed in this paper. In all MI based feature selection methods, effective and efficient estimation of high-dimensional MI is crucial. In this paper, a pruned Parzen window estimator and the quadratic mutual information (QMI) are combined to address this problem. The results show that the proposed approach can estimate the MI in an effective and efficient way. With this contribution, a novel feature selection method is developed to identify the salient features one by one. Also, the appropriate feature subsets for classification can be reliably estimated. The proposed methodology is thoroughly tested in four different classification applications in which the number of features ranged from less than 10 to over 15000. The presented results are very promising and corroborate the contribution of the proposed feature selection methodology.**

***Index Terms*—Feature selection, Parzen window estimator, quadratic mutual information (QMI), supervised data compression.**

I. Introduction

SEARCHING important features is essential to improve the accuracy, efficiency and scalability of a classification process [1]–[3]. This is especially important when one is required to deal with a large or even overwhelming feature set, such as a cDNA dataset [27]. In all feature selection methods, feature selection criteria are crucial. Different types of feature selection criteria divide feature selection methods into two categories: the filter method and the wrapper one [1], [3]. In the wrapper methods, the classification accuracy is employed to evaluate feature subsets, whereas, in the filter methods, various measurements may be used as feature selection criteria [1]. The wrapper methods may perform better, but huge computational effort is required [29]. Thus, it is difficult for them to deal with large feature sets, such as the gene (feature) set of a cDNA data. The feature selection criteria in the filter methods fall into two categories: the classifier-parameter-based criterion [4]–[6] and the classifier-free one [7]–[13]. This paper focuses on the later one. The popular classifier-free feature selection criteria are usually based on statistics computed from the empirical distribution. The first-order statistics were employed as the feature selection criteria in [7] and [8]. These feature selection criteria are sensitive to data noise [14]. Covariance, the second-order statistics, was employed as the feature selection criteria in [9] and [31]. The major drawback of these second-order statistics based criteria lies in their sensitivity to data transformation.

Recently, researchers have used mutual information (MI), which can be considered as higher order statistics [15], to identify the salient features [10]–[13], [16], [28]. The main advantages of MI are the robustness to noise and data transformation. This effect has been detailed in [10], [17], and [18]. Despite these advantages, the estimation of MI poses great difficulties as it requires the knowledge on the underlying probability density functions (pdfs) of the data space and the integration on these pdfs. In [10]–[13], and [16], histograms were used to estimate pdfs, and the computational difficulty of performing integration can be circumvented in a very efficient way. But the sparse data distribution, which is often encountered in a high-dimensional data space, may greatly degrade the reliability of histograms [16], [19], [20]. With these shortcomings, histograms are only applicable to a relatively low-dimensional data space. For example, in [12] and [16], feature selection methods were applied only to identify less than 10 features from the data sets with more than 1000 data patterns. In [10], [11], and [13], the problems about histograms were addressed. Only the two-dimensional (2-D) MIs were directly estimated, whereas the higher dimensional MIs were analyzed using the 2-D MI estimates. The experimental results previously presented in [10], [11], and [13] showed the effectiveness of these methods, but they are still inapt to direct estimate the high-dimensional MIs. This is quite a major shortcoming and causes two main problems to the MI based feature selection methods [10], [11], [13]. First, they are suboptimal in a sense that the selected features are individually, instead of globally, considered when the relevancy between the input features and the output class labels is measured. Thus, they are unable to deliver correct results when the relationship between the input and the output variables is strongly nonlinear. Second, without the knowledge on the amount of information contained in the selected features, the selection process cannot be stopped in an analytical way. Instead, the selection process must be terminated at a predetermined point. As a result, the selected features may either have certain useful information missed out or have included certain redundancy feature. The abovementioned problems can be overcome using direct estimation of the MI between the selected input features and the output. This is detailed in Section II-C.

In this paper, an effective and efficient approach to direct estimate MI is derived. A supervised data compression algorithm

Manuscript received Feburary 2, 2003; revised November 11, 2003. This work was supported by the Hong Kong SAR Government under CERG project 9040601-570.

The authors are with City University of Hong Kong, Hong Kong (e-mail: eetchow@cityu.edu.hk; dihuang@ee.cityu.edu.hk).

Digital Object Identifier 10.1109/TNN.2004.841414

is introduced for pruning a Parzen window estimator. Based on the pruned Parzen window estimator, quadratic mutual information (QMI) [18] is employed to reduce the computational complexity of MI estimation. We, accordingly, call the proposed MI estimate method as QMI-P. The presented results show that the proposed QMI-P can provide a huge computational reduction when one is handling a large data set. With the proposed QMI-P, a toward optimal feature selection methodology using MI (OFS-MI) is developed. In the proposed method, not only are the salient features effectively identified, the appropriate feature subsets can also be estimated in an analytical way. Two MI based criteria are used to guide the feature selection process. The first criterion is the feature relevance criterion (FRC). It searches for the important features. The second one is the feature similarity criterion (FSC), that is employed to avoid the highly redundant results. With these two criteria, the OFS-MI exhibits two major advantages. First, OFS-MI is able to determine the most prominent feature at each feature selection cycle. The direct MI estimation guarantees optimal or near-optimal features to be identified even when it is handling a highly nonlinear problem. At the same time, the introduced FSC avoids the high redundancy in the selected feature set. The effectiveness of OFS-MI can be clearly illustrated by the presented results. Second, attributed to the characteristics of the direct MI estimates, the feature selection process can be stopped analytically. Hence, the appropriate feature subsets can be systematically determined from any given large feature set. This contribution is important when a given data set contains a huge number of features, such as the cDNA data set consisting of over fifteen thousand features.

This paper is organized as follows. Section II gives the background of the MI-based feature selection method and summarizes the related work. Section III introduces the estimation of MI using QMI-P. Section IV details the proposed feature selection methodology, that consists of two MI-based criteria and a modified forward searching process. In Section V, it introduces several feature evaluation indexes. Then, extensive experimental results are presented in Section VI. And the conclusions are drawn in Section VII.

II. Background

A. Definition of MI

In accordance with Shannon's information theory [15], the uncertainty of a random variable C can be measured by the entropy $H(C)$. For two variables X and C, the conditional entropy $H(C|X)$ measures the uncertainty about C when X is known, and the MI $I(X;C)$ measures the certainty about C that is resolved by X. Apparently, the relation of $H(C)$, $H(C|X)$ and $I(X;C)$ is

$$\begin{aligned} H(C) &= H(C|x) + I(X;C), \text{ equivalently} \\ I(X;C) &= H(C) - H(C|X). \end{aligned} \tag{1}$$

The objective of training classification model is to minimize the uncertainty about predictions on class labels C for the known observations X. Thus, training a classifier is to increase the MI $I(X;C)$ as much as possible. Zero value of $I(X;C)$ means that the information contained in the observations (X) is not useful for determining their classes (C). The goal of a feature selection process for classification is naturally to achieve the higher values of $I(X;C)$ with the smallest possible size of feature subsets. With the entropy defined by Shannon, the prior entropy of C is expressed as

$$H_s(C) = -\sum_{c\in C} P(c)\log P(c) \tag{2}$$

where $P(c)$ represents the probability of C. The conditional entropy $H(C|X)$ is

$$H_s(C|X) = -\int_x p(x)\left(\sum_{c\in C} p(c|x)\log p(c|x)\right)dx. \tag{3}$$

The MI between X and C is

$$I_s(X;C) = \sum_{c\in C}\int_x p(c,x)\log\frac{p(c,x)}{P(c)p(x)}dx. \tag{4}$$

B. Estimation of MI

The pdfs are required to estimate the MI. One of the most popular ways to estimate MI is to use a histogram as a pdf estimator. In a 2-D data space, a histogram can be constructed feasibly, but there are problems in a high-dimensional data space. 1) The increase of data space dimension may have significantly degraded the estimation accuracy due to the sparse distribution of data, especially when the size of data set is relatively small [16], [19], [20]. 2) The required memories exponentially increase with the number of dimensions.

Continuous kernel-based density estimators can avoid the above problems. Actually, kernel-based density estimator is a more accurate approach than histograms [19], [20]. But with continuous kernel based pdf estimators, the integral operation in MI poses a major computational difficulty. In [18] and [22], the QMI (5) was developed

$$\begin{aligned} &I_{CS}(X_1;X_2) \\ &= \log\frac{\iint p_{12}(x_1,x_2)^2dx_1dx_2\iint(p_1(x_1)p_2(x_2))^2dx_1dx_2}{\left(\iint p_{12}(x_1,x_2)p_1(x_1)p_2(x_2)dx_1dx_2\right)^2}. \end{aligned} \tag{5}$$

When the Guassian function is employed for estimating the pdfs, the property of Gaussian function, i.e.

$$\int G\left(x-x_1,\sum\nolimits_1\right)G\left(x-x_2,\sum\nolimits_2\right)dx = G\left(x_1-x_2,\sum\nolimits_1+\sum\nolimits_2\right)$$

can simplify the integration in the QMI (5) to summation. In this way, the QMI (5) can be calculated efficiently. However, in a forward feature selection process where great deals of MI estimates are required, there is still a need to improve the computational efficiency of the MI estimation approach mentioned in [18] and [22], especially when one is handling a huge data set.

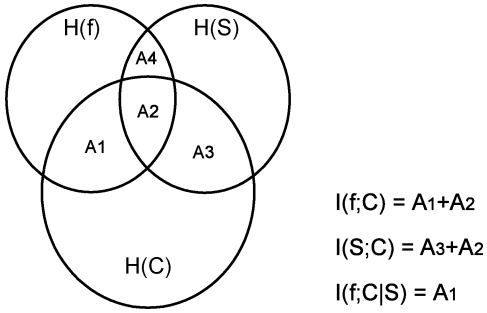


Fig. 1. Relation between input feature variables and output classes in terms of MI.

C. MI-Based Forward Feature Selection and the Related Work

The forward feature selection process in terms of MI [10], [11] can be formulized as follows:

1) (initialization) Set F to the initial feature set, S to the empty set;
2) $\forall f_i \in F$, compute $I(C; f_i)$;
3) find the feature f_k that maximizes $I(C; f_i)$, put f_k into S and delete it from F;
4) (greedy searching) repeat until the stopping criterion is met
a) calculate $I(C; S + f_i), \forall f_i \in F$;
b) choose the feature $f_k (f_k \in F)$ that maximizes $I(C; f_i + S)$. put f_k into S and delete it from F.
5) output the features set S.

Actually, in this forward process, the MI between the selected input variables and the output class labels $(I(S;C))$ gradually increases because the addition of input variables cannot decrease $I(S;C)$ [16]. Also, the incremental $I(S;C)$ decreases gradually to zero when all the relevant features are selected. Hence, this forward process can be stopped when the gradient of $I(S;C)$ is small enough when it indicates that the features beyond S contain negligable additional information for classification.

Due to the abovementioned high-dimensional histograms problems, Battiti [10] and Kwak [11] did not direct estimate the MI $I(f + S; C)$. In their methods, a high-dimensional conditional MI, $I(f; C|S)$, is analyzed with the related 2-D MI estimates. As shown in Fig. 1, $I(f + S; C)$ can be represented by the area $\mathrm{A}_1 + \mathrm{A}_2 + \mathrm{A}_3$, and $I(f; C|S)$ can be represented by A_1. To analyze $I(f; C|S)$, Battiti's MI feature selection (MIFS) used $I(f; C) - \beta \sum_{\dot{f} \in S} I(f; \dot{f})$, and Kwak's MI feature selection scheme under uniform information distribution (MIFS-U) employed $I(f; C) - \beta \sum_{\dot{f} \in S} I(f; \dot{f})(I(C; \dot{f})/H(\dot{f}))$.

These methods are acceptable because the area $\mathrm{A}_2 + \mathrm{A}_3$ is common for $f_i \in F$. But, the lack of direct estimation on the high-dimensional MI causes two main problems, i.e., suboptimization and the lack of estimation on the appropriate feature subsets. This has been detailed in Section I. The direct estimation on MI can overcome these two shortcomings. Hence, we consider a direct MI estimation. Also, it is worth noting that there is no principled guide to effectively deal with the redundancy of the given feature set in MIFS and MIFS-U. In these methods, the parameter β determines their capability of handling the redundancy and is quite important. But the principle on selecting the value of β was not discussed. The selection of β is, apparently, rather problem dependent. Different from the MIFS and the MIFS-U, the proposed methodology handles the highly redundant features in a principled way.

III. Estimation of MI Using the Proposed QMI-P

There is a dilemma when one is estimating MI with kernel-based density estimators. When the number of components in a kernel-based density estimator is large, estimating MI is computationally difficult or even impossible. On the other hand, it is desirable to have more components because more components can enhance the accuracy of the MI estimate. The requirements for correctly identifying the most prominent features, however, are not very strict in a sense that it is not necessary to determine the real MI. For a feature selection process, it is sufficient and necessary to guarantee that the relative ordering of the MI estimates remains correct. That is, given three variables X,Y,Z where it is known that $I(X;Y) > I(X;Z)$, the MI estimates are acceptable when the estimate of $I(X;Y)$ is larger than the estimate of $I(X;Z)$. Hence, it is possible that a relatively small data set is able to deliver accurate feature selection results. Thus, we employ a data compression procedure to reduce the computational complexity of the feature selection but without affecting the feature selection effectiveness. In this section, a data compression algorithm is proposed to prune a Parzen window estimator. Using the pruned Parzen window estimators and the QMI, the proposed estimation method (QMI-P) addresses the computational difficulty of estimating MI in an effective and efficient way. The contributions of QMI-P will be fully elaborated by the presented results.

A. Supervised Data Compression Algorithm

Two important issues must be taken into account when one is designing a data compression algorithm for estimating MI. First, a moderate compression assures a reliable MI estimation. Second, an unsupervised type of data compression algorithm is not preferred for estimating the MI between the input variables and output classes. In the proposed algorithm, data points are first partitioned into several clusters in a supervised way. Then, certain data points are sampled from the clusters in accordance with the size of the clusters. The supervised clustering algorithm described in this paper is an iterative process. At each iteration, the nearest cluster (u_m) is firstly determined from a randomly selected data point, x. If x has the same class label of u_m, it is put into u_m, or a new cluster is generated. The clustering process continues until all the data patterns are clustered. Let $X = (x_1, x_2, \ldots, x_{Nx})'$ and $C = (c_1, c_2, \ldots, c_{Nx})'$ be the input and the output data, respectively, M be the number of features, $L = \{l_1, l_2, \ldots, l_{Nl}\}$ be the class label set. The data compression algorithm can be expressed as follows.

```
Procedure supervised data compression
1. (Initialization)
 For class I_k(1 ≤ k ≤ N_l), create a cluster u_k in
which the center g_k is the average of the patterns
belonging to class l_k.
2. (Clustering)
 Repeat the following steps until there is no pat-
tern for clustering. Randomly select a data point
x. Find the nearest cluster
u_m to x according to the distance of x to the
cluster centers.
    . If x and u_m have the same class label, put x
into u_m,
and modify the center of u_m as the average of all
the data
points in the cluster u_m, otherwise, create a new
cluster for x. And x will be rejected in the re-
maining cluster process.
3. (Data sampling)
 Set the sampling data SX to be empty. Randomly
select certain
patterns from each cluster and put them into SX.
If the variance
of the data points in a cluster is less than σ_min,
put the one
pattern of that cluster into SX. For a cluster in
which the
variance is larger than σ_min, and if the number of
the patterns
of that cluster is more than 10, put 10 patterns
from that cluster
into SX, otherwise, put all the data patterns
into SX.
```

The parameter $\sigma_{\min}$ is selected as 0.05 because all the data used in this paper are normalized to have zero means and unit variances. Before further discussion, the following notations are briefed. Let the cluster set $U = \{u_1, u_2, \ldots, u_{Nu}\}$ be the result of the above supervised clustering process on $\{X; C\}$, n_j^x be the number of data points in cluster u_j. Thus, we have $N_x + = n_1^x + n_2^x + \ldots + n_{Nu}^x$. Let $SX = (sx_1, sx_2, \ldots, sx_{Ns})'$ be the results of the above data sampling process on $U = \{u_1, u_2, \ldots, u_{Nu}\}$, and n_j^s be the number of the data points sampled from cluster u_j. The size of SX is $N_s = n_1^s + n_2^s + \ldots + n_{Nu}^s$. With these notations, the proposed compression algorithm can be described as a process in which all the data points in X are firstly partitioned into N_u homogenous subgroups. Then, the number of data points in each subgroup (say u_j) is reduced from n_j^x to n_j^s.

B. Proposed Method to Estimate MI—QMI-P

Conventional Parzen window estimator [21] assumes that X is uniformly distributed, i.e., $p(x_i) = 1/Nx$. In the proposed density estimator, in order to make use of more information contained in the original data set, the assumption of a uniform distribution is not applied. Based on the compressed data set SX, we calculate $p(sx_i)$ $(sx_i \in u_j)$ with

$$p(sx_i) = \left(\frac{n_j^x}{Nx}\right) \times \left(\frac{1}{n_j^s}\right). \tag{6}$$

Hence, with the compressed data set SX, the marginal and the conditional pdf can be estimated as

$$\begin{aligned} p(x) &= \sum_{i=1}^{Ns} p(sx_i) p(x|sx_i) \\ &= \sum_{i=1}^{Ns} p(sx_i) \kappa(x - sx_i, \Sigma_i) \qquad (7) \\ p(x|l) &= \sum_{sx_i \in class\, l} p(sx_i) p(x|sx_i) \\ &= \sum_{sx_i \in class\, l} p(sx_i) \kappa(x - sx_i, \Sigma_i) \qquad (8) \\ p(l) &= \sum_{sx_i \in class\, l} p(sx_i), \qquad (9) \end{aligned}$$

where

$$\begin{aligned} \kappa(x - x_0, \Sigma) &= G(x - x_0, \Sigma_0) \\ &= \frac{1}{(2\pi h)^{M/2} |\Sigma_0|^{1/2}} \exp\left(-\frac{(z - z_0)^T \Sigma_0^{-1} (z - z_0)}{2h^2}\right). \end{aligned} \tag{10}$$

In the Gaussian function (10), Σ_0 is determined from the variance matrix of the overall data, h is the bandwidth of the kernel function. In this paper, all input data are normalized to unit variances. Hence, we set $\Sigma_0 = I$. And, the method in [23] is used to determine h, i.e.,$h = (4/(M+2))^{1/(M+4)} Nx^{-1/(M+4)}$. With the pdf estimators shown in (7)–(9), the QMI can be calculated as

$$I_{CS}(X; C) = \log\left(\frac{V_{(c,x)^2} V_{(c)^2} V_{(x)^2}}{\left(V_{(cx)}^2\right)}\right) \tag{11}$$

where

$$\begin{aligned} V_{(c,x)^2} &= \sum_l \int p(l, x)^2 dx \\ &= \sum_{k=1}^{Nl} \sum_{sx_i \in \text{class } l_k} \sum_{sx_j \in \text{class } l_k} p(sx_i) p(sx_j) \\ &\quad \times G(sx_i - sx_j, 2I) \\ V_{(c)^2} &= \sum_{k=1}^{Nl} p(l_k)^2 \\ &= \sum_{k=1}^{Nl} \left(\sum_{sx_i \in \text{class } l_k} p(sx_i)\right)^2 \\ V_{(x)^2} &= \int p(x)^2 dx \\ &= \sum_{i=1}^{Ns} \sum_{j=1}^{Ns} p(sx_i) p(sx_j) G(sx_i - sx_j, 2I), \end{aligned}$$

$$V_{(cx)} = \sum_{l} \int p(l,x)p(l)p(x)dx$$
$$= \sum_{k=1}^{Nl} \left(\sum_{sx_i \in \text{class } l_k} p(sx_i) \right) \cdot \sum_{j=1}^{Ns} \sum_{sx_i \in \text{class } l_k} p(sx_j)p(sx_i)G(sx_j - sx_i, 2I).$$

C. Evaluation of QMI-P

In order to evaluate the proposed QMI-P, comparisons were made by using a synthetic data set consisting of two classes {0, 1}. The data were generated from a mixture of two normal distributions

$$\text{Class 0 (n data points)}\ X \sim N\left([0.5 \quad 0.5], \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}\right)$$
$$\text{Class 1 (n data points)}\ X \sim N\left(-[0.5, -0.5], \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}\right).$$

The overlapping between the two classes is determined by the variance σ^2. It can be concluded that a small overlapping between the two classes means a larger value of MI $I(X;C)$, i.e., $I(X;C)$ increases with the decrease of σ^2. In our study, three estimates of MI were compared. MI-p represents the result calculated using the proposed QMI-P; MI-w represents the result calculated using the conventional Parzen window estimators constructed with the whole data set $\{X;C\}$; and MI-s represents the results of the conventional Parzen window estimators constructed with the compressed data set SX. In this section, the real MI (MI-r) is used as reference. The numerical analysis method was employed to approximate MI-r, i.e.

$$I(X;C) = \sum_{c=0,1} \int p(x,c) \log \frac{p(x,c)}{p(x)p(c)} dx \approx \sum_{c=0,1} \sum_{x_i} p(x_i,c) \log \frac{p(x_i,c)}{p(x_i)p(c)} \Delta x_i. \quad (12)$$

61×61 data samples evenly distributed in the range of $(-3,-3)$–$(3, 3)$ were used to calculate (12).

Given a fixed size of data set, say 2000 data patterns, different values of the variance σ^2 were tried. For the same σ^2, 10 independently data sets were generated and tested. The averages of these 10 trials are presented in Fig. 2. The relative ordering of the MI values can be correctly estimated by using MI-p and MI-w. Based on the compressed data set, MI-s is found to be an unreliable approach mainly because too much information of the original data set was discarded when estimating the MI.

MI-p is further compared with MI-w in terms of computational efficiency. The computational complexity of MI-p and MI-w mainly depends on the number of components of the pdf estimators. With an increase on the size of data set, the computational complexity of MI-w increases rapidly, whereas the computational complexity of MI-w might not be significantly increased with the contribution of the proposed data compression algorithm. In our study, with $\sigma^2 = 0.5$, the data sets of different sizes (from 400 to 5800 data patterns) were used to give comparisons of the running time (t_{MI}) for calculating MI-p and MI-w. Fig. 3 illustrates these comparative

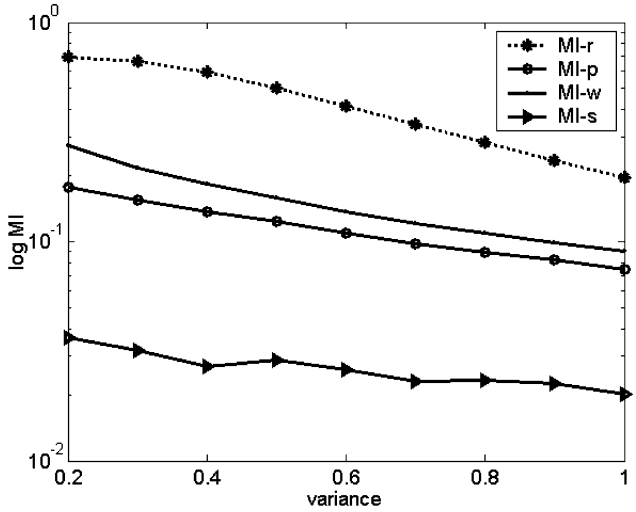


Fig. 2. Comparison of different MI estimates. The presented results are the averages of 10 trials. MI-p is the result of the proposed approach QMI-P; MI-w and MI-s is the result of the approach proposed in [18] on the original data and the reduced data, respectively.

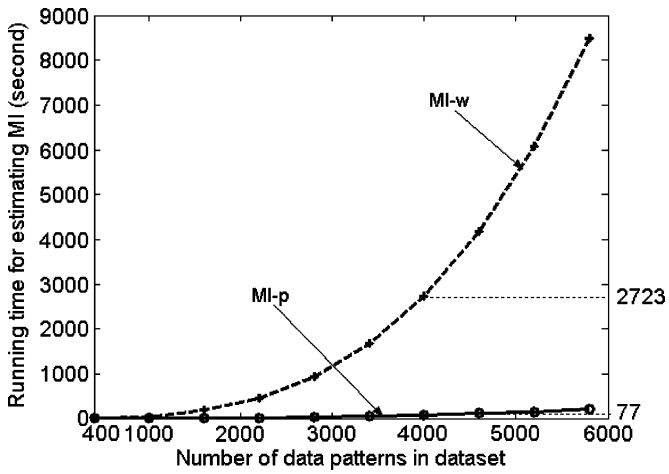


Fig. 3. Comparison of running time t_{MI} between MI-p and MI-w. MI-p is the results of the proposed approach QMI-P; MI-w is the result of the approach proposed in [18].

results. Obviously, with the increase of the number of data patterns, the computational saving caused by the data compression algorithm becomes increasingly significant. In the example of 4000 data patterns, t_{MI} for calculating MI-w is about 40 times of that for calculating MI-p.

The above comparisons in terms of their correctness and the computational efficiency show that the proposed QMI-P is the most suitable approach for a MI-based forward feature selection process.

IV. Towards Optimal Feature Selection Using QMI-P

The proposed OFS-MI consists of two MI-based criteria, namely FRC and FSC, and a modified forward searching algorithm.

A. MI-Based Criteria

The feature relevancy criterion FRC is aimed at selecting the relevant features. And the FSC is used to reduce the redundancy

in the selection results. Both of them are calculated by using the proposed estimation approach QMI-P.

Suppose that a feature set S has been selected from $\{X;C\}$. The FRC of a feature $f_m(f_m+ \notin \mathrm{S})$ is defined as

$$\mathrm{FRC}(f_m) = I(S + f_m; C) \tag{13}$$

which is calculated by using (11).

The FSC is to measure the similarity between the feature subset S and a single feature $f_m(f_m \notin \mathrm{S})$. It is known that the MI $I(S; f_m)$ must be large when the MI of f_m with any feature in S is large. That is, if f_m is very similar to $f_i(f_i \in S)$, f_m must be similar (or redundant) to S. With this idea, the FSC of a feature $f_m(f_m \notin S)$ to a feature set S, is defined as

$$\mathrm{FSC}(f_m) = \arg\max_{f_i \in S} \left(\frac{I(f_m; f_i)}{H(f_i)} \right). \tag{14}$$

Based on the relationship between entropy and the MI described in (1), we have

$$H(f_i) = H(f_i|f_m) + I(f_i; f_m) \geq I(f_i; f_m).$$

Obviously, $\mathrm{FSC}(f_m) \leq 1$ with equality if and only if f_m has no additional information beyond S. When $\mathrm{FSC}(f_m)$ is large enough, i.e., $\mathrm{FSC}(f_m) \geq \theta$, the feature f_m can be considered as a redundant feature for S, and should not be added into S. Throughout the paper θ is set to 0.9.

Only the 2-D MI estimates are needed for calculating FSC (14). For the features f_m and f_j, the results obtained in Section III-A is $SX = \{(sx_{m1}, sx_{j1}), (sx_{m2}, sx_{j2}), \ldots\ldots, (sx_{mN_s}, sx_{jN_s})\}$. $I(f_j; f_m)$ in (14) can be estimated by using

$$I(f_j; f_m) = \log \frac{V_{(f_j,f_m)^2} V_{(f_j)^2} V_{(f_m)^2}}{V^2_{(f_{jm})}} \tag{15}$$

where

$$V_{(f_j,f_m)^2} = \sum_{k=1}^{Ns} \sum_{i=1}^{Ns} \prod_{q=j,m} p(sx_k) p(sx_i) G(sx_{qk} - sx_{qi}, 2I)$$

$$V_{(f_q)^2} = \sum_{k=1}^{Ns} p(sx_i) \sum_{i=1}^{Ns} p(sx_k) G(sx_{qk} - sx_{qi}, 2I)$$

$$V_{(f_{jm})} = \sum_{k=1}^{Ns} p(sx_k) \prod_{q=j,m} \left(\sum_{i=1}^{Ns} p(sx_i) G(sx_{qk} - sx_{qi}, 2I) \right).$$
$$q = j, m.$$

The entropy $H(f_i)$ in (14) can also be estimated by using (15) in that $H(f_i) = I(f_i; f_i)$.

B. Forward Feature Selection Process With the MI-Based Criteria

The proposed OFS-MI is a forward and iterative process, which begins with an empty feature set and additional features are included one by one. Based on FRC and FSC, the forward feature selection process in the OFS-MI is realized as follows. The diagram of OFS-MI is shown in Fig. 4.

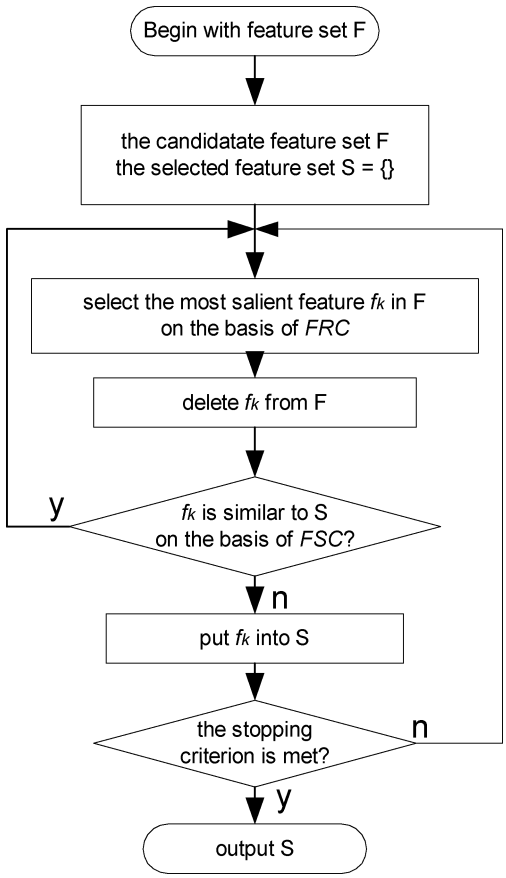


Fig. 4. Block diagram of the proposed feature selection methodology OSF-MI.

Step 1) Set F to the initial feature set, S to the empty set, the iteration number $j = 0$.

Step 2) Find out the feature f_k having the maximal FRC. Put f_k into S and delete it from F. The iteration number j is 1. Remember the FRC of this iteration as $\mathrm{FRC}_j = \mathrm{FRC}(f_k)$.

Step 3) Estimate $\mathrm{FRC}(f_i) = I(C; S + f_i), \forall f_i + \in F$.

Step 4) Repeat until an appropriate feature is added into S or there is no features remaining in F.

a) Choose the feature f_k having the maximal FRC, and delete f_k from F.

b) If $\mathrm{FSC}(f_k) < \theta$, then put f_k into S, $j = j + 1$, $\mathrm{FRC}_j = \mathrm{FRC}(f_k)$.

Step 5) If $(\mathrm{FRC}_j - \mathrm{FRC}_{j-1})/\mathrm{FRC}_1 \leq \gamma$, then delete f_k from S and goto Step 6, otherwise, goto Step 3.

Step 6) Estimate the appropriate number of selected features ANSF as $B \leq \mathrm{ANSF} \leq \mathrm{cardinality}(S)$, where B satisfies $\mathrm{FRC}_B \geq \lambda \times \max(\mathrm{FRC}_i)$.

Step 7) Output the set S and ANSF.

V. Feature Subset Evaluation Index

In this paper, four indexes, namely NNG, KNNG, *cs*, and *scbd*, are adopted to evaluate the effectiveness of the selected feature subsets. The first two indexes are based on the results of different classification models.

A. NNG and KNNG

NNG represents generalization accuracy of a standard neural network based classifier, KNNG represents that of the k-NN rule. All these classifier models are available in Weka software [30]. The parameters of classifiers were set to the default values in our study. Obviously, among the several feature subsets of the same size, the one with the best classification performance is considered to be the best.

B. Class Separability

Class separability (cs) of a data set is calculated by

$$cs = \text{trace}\left(S_w^{-1} S_b\right) \tag{16}$$

where S_b is the between class scatter matrix, and S_w is the within class scatter matrix [24]. Assume that μ_i is the mean vector of the data points belonging to the class c_i, μ is the mean vector of all the data points, and π_i is the *a priori* probability that a data point belongs to class c_i, we have

$$S_b = \sum_{i=1}^{N_c} \pi_i (\mu_i - \mu)(\mu_i - \mu)^T$$

$$\text{and} \quad S_w = \sum_{i=1}^{N_c} \pi_i \sum_{x \in c_i} (x - \mu_i)(x - \mu_i)^T$$

where N_c is the number of class. A high value of the class separability index ensures that the classes are well separated by their scatter means. That is, in order to obtain a better classification performance, a large value of cs is preferred.

C. Square Chernoff Bound Distance

In this paper, a new feature evaluation index, square chernoff bound distance $(scbd)$, is introduced. For the two classes l_i, l_j, the minimum classification error attainable by Bayes classifier p_b [25] satisfies

$$p_b = \int_{-\infty}^{+\infty} \min\left(P(l_i)p(x|l_i), P(l_j)p(x|l_j)\right) dx$$

$$\leq P(l_i)^s P(l_j)^{1-s} \int_{-\infty}^{+\infty} p(x|l_i)^s p(x|l_j)^{1-s} dx$$

where $0 \leq s \leq 1$. Based on this inequality, for the classification data set $\{X; C\}$ mentioned in Section III-A, $scbd$ is defined as

$$scbd = \sum_{i=1}^{Nl} \sum_{j=1}^{i} scbd_{ij}$$

$$= \sum_{i=1}^{Nl} \sum_{j=1}^{i-1} p(l_i)p(l_j) \int_{-\infty}^{+\infty} \left(p(x|l_i)p(x|l_j)\right) dx. \tag{17}$$

Obviously, $scbd_{ij}$ can be calculated by putting

$$scbd_{ij} = P(l_i)P(l_j) \sum_{x_m \in class\, l_i} \sum_{x_n \in class\, l_j} G(x_m - x_n, 2I).$$

A feature subset with a small $scbd$ is able to deliver better classification results.

VI. Results and Discussion

In this section, the proposed OFS-MI is evaluated from two main perspectives—the effectiveness on the feature selection, and the reliability of estimating the appropriate number of the ANSF. Despite the fact that MIFS and MIFS-U are not able to estimate the ANSF, they are included in our comparative study to highlight the improvement caused by the direct MI estimates at the high-dimensional scenario. The comparative results are expressed in terms of the evaluation indexes mentioned in Section V. As for verifying the estimates on ANSF, *a priori* knowledge in a synthetic data set can be used. In a real application, verifying the estimate on ANSF poses practical difficulties because the ideal ANSF is not known. In this study, we use the performance of different classifiers to handle this difficulty. When all the classifiers can obtain their best or near-best performance within the ANSF, it can be reasonably concluded that the estimation of the ANSF is reliable. Two types of classifiers, the k-NN rule and a standard neural network model, are employed for this purpose.

In this study, four classification data sets were used. These include two synthetic data sets and two real data sets. The first synthetic data set highlights the advantages of the proposed OFS-MI when dealing with highly nonlinear problem. The highly redundant LED domain data set was used to verify the capability of dealing with the redundancy and irrelevance in the feature set. And the cDNA ovarian cancer classification data is used to highlight the importance of introducing the OFS-MI when one is required to deal with huge number of features.

In the OFS-MI, λ $(0 < \lambda < 1)$ is used to determine the range of the number of selected features. As shown in Fig. 5 and Fig. 7, too small values of λ will lead to a wider range of ANSF. In this case, the feature sets with little information are considered appropriate ones. Clearly, without enough information, these feature sets could not assure a respectable classification performance. Thus, a relative large λ is preferred. Throughout this study, $\lambda = 0.8$ is used. Also, γ is used to determine when the selection process should be stopped. Ideally, γ should be as close to zero as possible in order to include all the required information. In this paper, $\gamma = 0.01$ is considered a small enough index for delivering a promising performance.

A. Synthetic Highly Nonlinear Data Set

The four-dimension (4-D) synthetic data $X = \{X_1, X_2, X_3, X_4\}$ was generated in this experiment, which consists of 2000 patterns belonging to two classes $\{1, -1\}$. For the input variable X_1, 500 patterns were generated from the normal distribution $\aleph(-0.4, 0.3)$, and other 1500 patterns were generated from $\aleph(0.2, 0.3)$; for X_2, 500 patterns were generated from the normal distribution $\aleph(-0.1, 0.3)$, and other 1500 patterns were generated from $\aleph(0.2, 0.3)$. The input variable X_3 is equal to the square of X_1, and X_4 is equal to the square of X_2. The class of a pattern $x = \{x_1, x_2, x_3, x_4\}$ is determined by $c = sign(x_3 - x_4)$. Obviously, the input

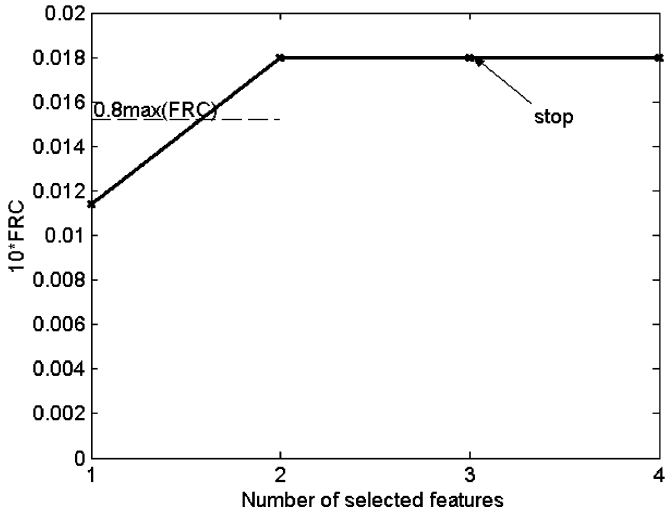


Fig. 5. FRC on a highly nonlinear data set.

TABLE I
ORDER ON SELECTION OF A HIGHLY NONLINEAR DATASET WHICH CONSISTS OF THE FOUR INPUT VARIABLES X_1, X_2, X_3, X_4

β	0	0.3	0.7	1
MIFS	$\{X_1, X_3, X_2, X_4\}$	$\{X_1, X_2, X_3, X_4\}$	$\{X_1, X_2, X_3, X_4\}$	$\{X_1, X_2, X_3, X_4\}$
MIFS-U		$\{X_1, X_2, X_3, X_4\}$	$\{X_1, X_2, X_3, X_4\}$	$\{X_1, X_2, X_3, X_4\}$
OFS-MI	$\{X_3, X_4, X_2, X_1\}$			

variables X_3 and X_4 are considered more important than X_1 and X_2 in this classification task. The proposed data compression method greatly reduced the number of data patterns from 2000 to 99.

Table I shows the selection results of the compared methods. It indicates that only OFS-MI is able to identify the relevant features correctly. Both MIFS and MIFS-U are not able to obtain the correct results when they are handling this highly nonlinear classification problem. In Fig. 5, which shows the change of FRC with the number of selected features, it shows that ANSF $= 2$ is estimated. This result is consistent with the *a priori* knowledge of this data set.

B. Highly Redundant LED Data Set

The data set of LED display domain [26] has 24 features, in which the first 7 features determine the class label of a pattern, whilst the rest 17 features are irrelevant. In our study, 1000 data patterns were generated. In order to further examine the capabilities of dealing with irrelevant features and redundant features, other 24 redundant features are added to increase the total number of features to 48. These 24 redundant features are generated from corrupting the original 24 features by noise with a normal distribution $\aleph(0, 0.01)$. After the proposed data compression process, the number of data patterns was reduced from 1000 to about 500.

First, OFS-MI was compared with others in terms of their effectiveness. With the *a priori* knowledge, all the compared feature selection methods were required to select 7 features, and their selection results were listed in Table II. In this table, f_i and f_i' denote the original feature and its corresponding redundant feature, respectively. The redundant features were underlined in Table II. Apparently, when we were dealing with this highly redundant feature set, only MIFS with $\beta = 0.5$ or $\beta = 1$ and OFS-MI are able to deliver the correct results, whereas MIFS-U failed to avoid the redundancy in the selected feature set in all cases. Also, it shows that β is very important to MIFS and MIFS-U in a way that it can greatly affect their selection results. But as analyzed in Section II-C, there is not discussion on selecting the value of β.

TABLE II
RESULTS OF FEATURE SELECTION METHODS ON THE HIGHLY REDUNDANT LED DOMAIN DATA SET. THE REDUNDANT FEATURES ARE UNDERLINED

Feature selection method	Order of feature selection
MIFS-U $\beta = 0$	$f_5', f_2', \underline{f_5}, \underline{f_2}, f_7', f_4, f_7$
MIFS $\beta = 0.5$	$f_5', f_2', f_4', f_7', f_3', f_1, f_6$
MIFS $\beta = 1$	$f_5', f_2', f_4, f_1, f_3, f_7, f_6$
MIFS-U $\beta = 0.5$	$f_5', f_2', f_4', f_7', \underline{f_5}, \underline{f_2}, f_3'$
MIFS-U $\beta = 1$	$f_5', f_2', f_4', f_7', f_3', f_1', \underline{f_5}$
OFS-MI	$f_5', f_2, f_4, f_3, f_7, f_1, f_6$

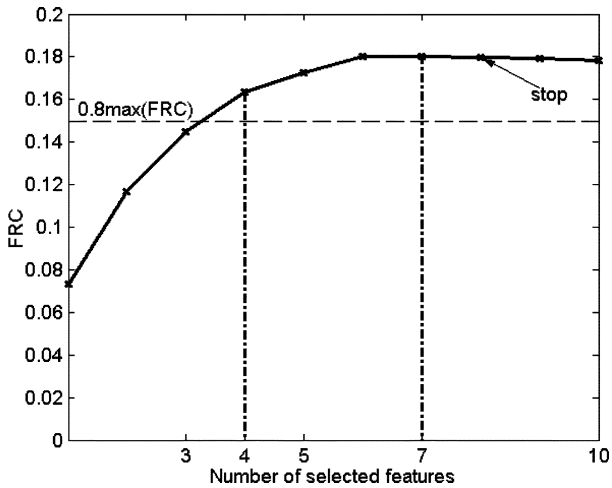


Fig. 6. FRC on the highly redundant LED data set.

Second, the change of FRC with the number of selected feature is shown in Fig. 6. It shows that an estimate of $4 \leq$ ANSF ≤ 7 is obtained. And based on the knowledge that only the first 7 features are relevant to classification task, it can be concluded that the estimate of $4 \leq$ ANSF ≤ 7 is correct.

C. Spambase Data Set

This data set was used to classify emails into spam or non-spam category. It consists of 4601 patterns in which 2000 are training patterns, and other 2601 are testing patterns. It has 57 numerical input variables and 2 output classes.

The effectiveness of the different feature selection methods was compared. All these methods were required to select 3, 6, 9, 12, 15, 18, 21, and 24 features, respectively. Fig. 7 illustrates the comparative results in terms of four evaluation indexes, in which (a) and (b) indicate that the proposed OFS-MI enables different classifiers to achieve the best performance in most cases. Also, the performance of the OFS-MI is much better especially when the number of the selected features is less than 21. In terms of

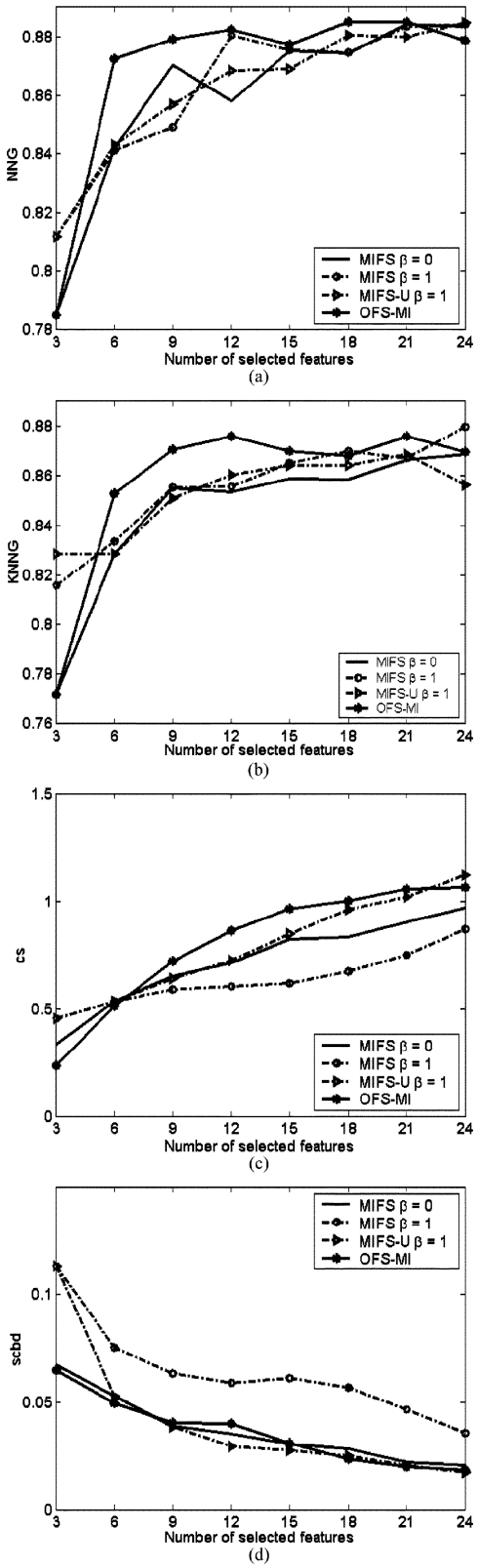


Fig. 7. Comparisons of different feature selection methods on the spambase dataset. (a) NNG: the generalization accuracy of neural network based classifier. (b) KNNG: the generalization accuracy of k-NN rule. (c) cs: class separability. (d) $scbd$: square chernoff bound distance.

other evaluation indexes, cs and $scbd$, the proposed OFS-MI is still able to deliver better results than other methods, as shown in Fig. 7(c) and (d).

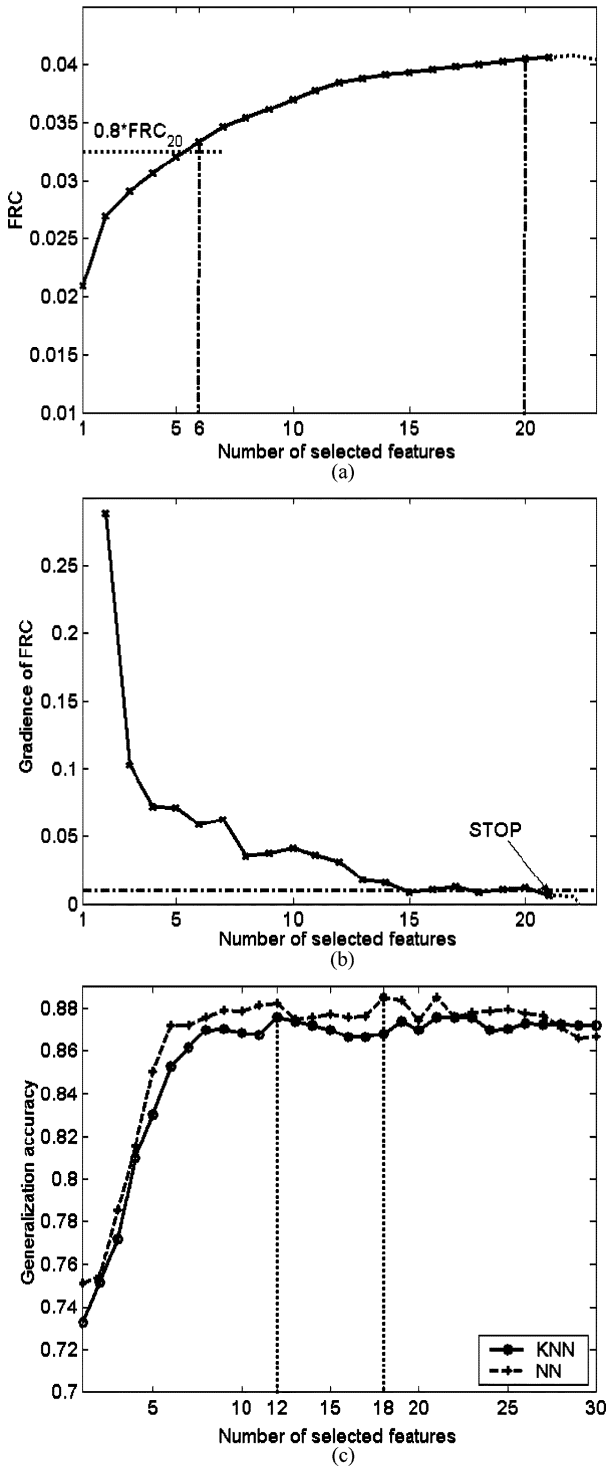


Fig. 8. FRC and its gradient on the spambase dataset. (a) FRC. (b) Gradient of FRC. (c) Classification results to verify the estimate on ANSF.

In this application, the changes of FRC and its gradient with the number of selected features are illustrated in Fig. 8(a) and (b). The feature selection process in OFS-MI stopped when 21 features were selected, and it was estimated that $6 \leq \text{ANSF} \leq 20$. The classification performance of the k-NN and the neural network model are illustrated in Fig. 8(c), which verifies the estimate of $6 \leq \text{ANSF} \leq 20$. It is noted that the k-NN and the neural network model achieved their best performance when the numbers of selected features were 12 and 18, respectively.

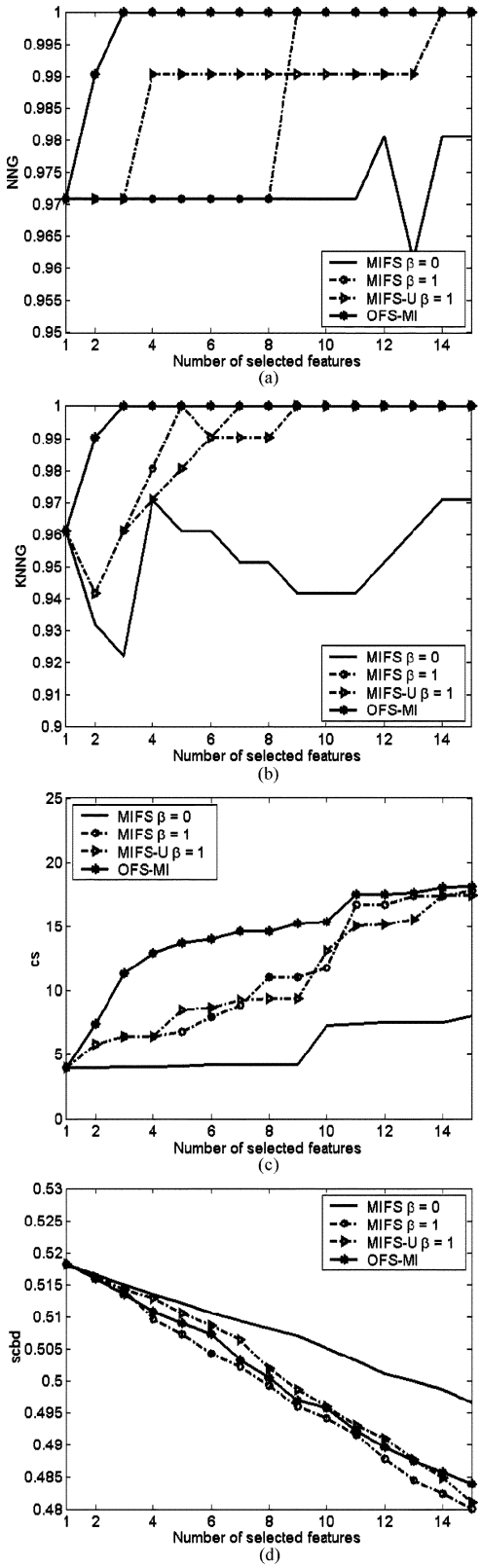


Fig. 9. Comparisons of different feature selection methods on the ovarian classification cDNA dataset. (a) NNG: NN-based classifier generalization accuracy. (b) KNNG: k-NN rule generalization accuracy. (c) cs: class separability. (d) $scbd$: square chernoff bound distance.

All these classifiers were able to obtain the best or the near best performance within $6 \leq$ ANSF ≤ 20. The estimate of $6 \leq$ ANSF ≤ 20 is reliable.

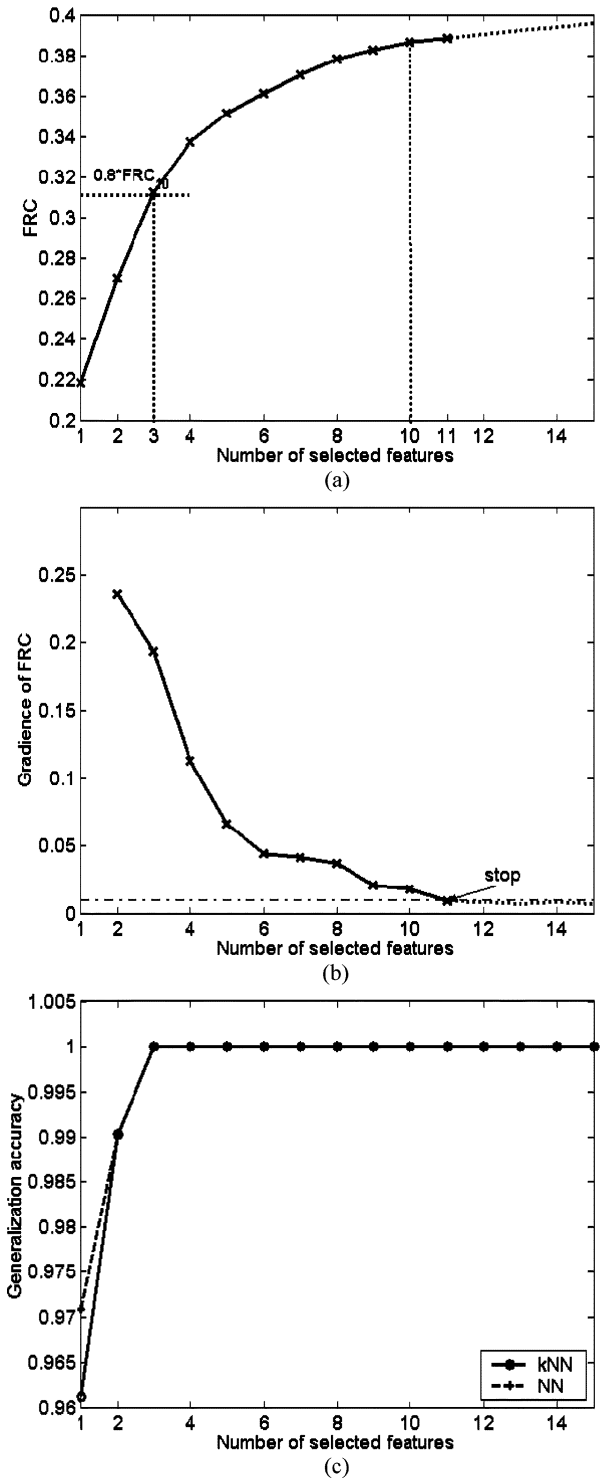


Fig. 10. FRC and its gradient on the ovarian classification cDNA data set. (a) FRC. (b) Gradient of FRC. (c) Classification results to verify the estimate on ANSF.

D. cDNA Data for Ovarian Cancer Classification

The objective of this classification is to identify proteomic patterns in serum that distinguishes ovarian cancer from non-cancer [27]. There are 253 data samples, and each sample contains 15154 features—genes. We partitioned the data set into

two disjoint parts, 150 data samples were used for training and 103 were used for testing. The proposed data compression approach was not applied to this data set because the original data set is small. Hence, all the data points were used to construct the pdf estimators, i.e., $SX = X$.

In order to effectively dealing with the data set containing a large quantity of features, all the compared methods filtered out the features with no or less information before the feature selection process begins to run. That is, all the features were firstly ranked in the order of MI $I(f;C)$. Then, only the most important 600 features were left behind for the forthcoming iterative feature selection process.

In Fig. 9, the comparison results of OFS-MI, MIFS, and the MIFS-U are illustrated in Fig. 9. Obviously, OFS-MI is much more effective than the other methods in this example. The changes of FRC and its gradient with the number of selected features are shown in Fig. 10(a) and (b). With these results, OFS-MI delivered a result of $3 \leq \text{ANSF} \leq 10$. In Fig. 10(c), it illustrates the classification accuracy of the k-NN and the neural network model. It is noted that, using the features selected by OFS-MI, the k-NN and the neural network model can deliver their best performance with the first three features. Accordingly, the estimate of $3 \leq \text{ANSF} \leq 10$ is considered reliable in this application.

VII. Conclusion

This paper describes a newly developed MI-based feature selection methodology. Attributed to the MI-based criteria, the proposed feature selection methodology offers two major contributions. First, the optimal or the near-optimal features can be effectively identified. The problem of highly redundant features is also addressed. Second, the appropriate feature subsets are estimated in a systematic way. The later contribution is important when one is dealing with a huge feature set, such as the cDNA data set. The generalization performance or the *a priori* knowledge corroborates that the proposed methodology is able to provide reliable estimates on the appropriate feature subsets. This type of information is essential for constructing an appropriate classifier.

Acknowledgment

The authors would like to thank the editors and the anonymous reviewers for their useful comments.

References

[1] J. W. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Mateo, CA: Morgan Kaufmann Publishers, 2001, pp. 116–121.

[2] R. Kohavi and G. H. John, "Wrapper for feature subset selection," *Artif. Intell.*, vol. 97, pp. 273–324, Dec. 1997.

[3] E. P. Xing, M. I. Jordan, and R. M. Krap, "Feature selection for high-dimensional genomic microarray data," in *Proc. 18th Int. Conf. Machine Learning (ICML)*, 2001, pp. 601–608.

[4] W. Wang, P. Jones, and D. Partridge, "A comparative study of feature-salience ranking techniques," *Neural Computat.*, vol. 13, no. 7, pp. 1603–1623, 2001.

[5] R. Setiono and H. Liu, "Neural-network feature selector," *IEEE. Trans. Neural Netw.*, vol. 8, no. 3, pp. 654–661, May 1997.

[6] J. Utans and J. Moody, "Selecting neural network architectures via the prediction risk: application to corporate bond rating prediction," in *Proc. 1st Int. Conf. Artificial Intelligence Applications on Wall Street*, Los Alamitos, CA, 1991, pp. 4–35.

[7] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature selection," *IEEE. Trans. Comput.*, vol. C-26, no. 9, pp. 917–922, Sep. 1977.

[8] H. Lin, H. Motoda, and M. Dash, "A monotonic measure for optimal feature selection," in *Proc. 10th Eur. Conf. Machine Learning*, Chemnitz, Germany, 1998, pp. 101–106.

[9] E. Permot and F. Vallet, "Determining the relevant parameters for the classification on a multi-layer perceptron: application to radar data," in *Proc. Int. Conf. Artificial Neural Networks*, Espoo, Finland, 1991, pp. 797–802.

[10] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

[11] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," *IEEE Trans. Neural Netw.*, vol. 13, no. 1, pp. 143–159, Jan. 2002.

[12] B. V. Bonnlander and A. S. Weigend, "Selecting input variables using mutual information and nonparametric density estimation," in *Proc. Int. Symp. Artificial Neural Network (ISANN)*, Taiwan, 1994, pp. 42–50.

[13] A. Al-ani and M. Deriche, "An optimal feature selection technique using the concept of mutual information," in *Proc. 6th Int. Symp. Signal Processing and Its Applications (ISSPA)*, Kuala Lumpar, Malaysia, 2001, pp. 477–480.

[14] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, no. 3, pp. 131–156, May 1997.

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[16] H. H. Yang and J. Moody, "Feature selection based on joint mutual information," in *Proc. Int. ICSC Symp. Advances in Intelligent Data Analysis*, Rochester, New York, 1999, pp. 342–349.

[17] W. T. Li, "Mutual information functions versus correlation functions," *Stat. Phys.*, vol. 60, no. 5/6, pp. 823–837, Sep. 1990.

[18] J. C. Principe, D. Xu, and J. Fisher III, "Information theoretic learning," in *Unsupervised Adaptive Filtering*, S. Haykin, Ed. New York: Wiley, 2000, pp. 265–320.

[19] M. Young, B. Rajagopalan, and U. Lall, "Estimation of mutual information using kernel density estimators," *Phys. Rev. E*, vol. 52, no. 3(B), pp. 2318–2321, Sep. 1995.

[20] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Phys. Rev. A*, vol. 33, no. 2, pp. 1134–1140, Feb. 1986.

[21] E. Parzen, "On the estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, pp. 1064–1076, 1962.

[22] K. Torkkola and W. M. Campbell, "Mutual information in learning feature transformations," in *Proc. 17th ICML*, Stanford, CA, 2000, pp. 1015–1022.

[23] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Chapman & Hall, 1986.

[24] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1982.

[25] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. London, U.K.: Academic, 1999, pp. 153–154.

[26] UCI machine learning repository.. [Online] Available http://www.ics.uci.edu/~mlearn/MLRepository.html

[27] Available http://clinicalproteomics.steem.com.. [Online]

[28] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1667–1671, Dec. 2002.

[29] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Proc. 11th Int. Conf. Machine Learning (ICML)*, San Francisco, CA, 1994, pp. 121–129.

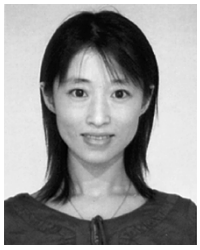
[30] Available http://www.cs.waikato.ac.nz/~ml/weka.. [Online]

[31] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Dept. Comput. Sci., Waikato Univ., New Zealand, 1999.

Tommy W. S. Chow (M'94–SM'03) received the B.Sc. (First Hons.) and Ph.D. degrees from the University of Sunderland, Sunderland, U.K.

In 1988, he joined the City University of Hong Kong, Hong Kong, as a Lecturer. He is currently a Professor in the Department of Electronic Engineering. His research has been in the areas of system identification, machine fault diagnostics, and neural network theory and applications. He has authored or coauthored of one book chapter and over 90 journal articles related to his research.

Prof. Chow was the Chairman of the Control Automation and Instrumentation Division, Hong Kong Institute of Engineers, during 1997–1998.



D. Huang received the B.E. degree from the Liaoning Institute of Technology, Jinzhou, P.R.C, and the M.E. degree from the Shenyang Institute of Automatics, Chinese Academy of Science, Shenyang, P.R.China She is currently working toward the Ph.D. degree in the Department of Electronic Engineering, City University of Hong Kong.

Her research interest areas are neural networks, pattern recognition, and their applications.

Ms. Huang obtained the Best Paper award at the 2002 IEEE Industrial Electronics Society Annual Meeting in Seville, Spain.