Efficiently Searching the Important Input Variables Using Bayesian Discriminant

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

***Abstract*—This paper focuses on enhancing feature selection (FS) performance on a classification data set. First, a novel FS criterion using the concept of Bayesian discriminant is introduced. The proposed criterion is able to measure the classification ability of a feature set (or, a combination of the weighted features) in a direct way. This guarantees excellent FS results. Second, FS is conducted by optimizing the newly derived criterion in a continuous space instead of by heuristically searching features in a discrete feature space. Using this optimizing strategy, FS efficiency can be significantly improved. In this study, the proposed supervised FS scheme is compared with other related methods on different classification problems in which the number of features ranges from 33 to over 12,000. The presented results are very promising and corroborate the contributions of this study.**

***Index Terms*—*A posteriori* probability, Bayesian discriminant (BD), feature selection (FS), Parzen window estimator.**

I. Introduction

FEATURE selection (FS) is an essential preprocessing to guarantee high accuracy, efficiency, and scalability of classification [1], especially when one is dealing with a huge data set in the areas such as image processing and bioinformatics. Generally, an FS scheme consists of two main parts—the criterion for evaluating the "goodness" of features and the feature searching engine to optimize certain evaluation criterion.

Obviously, the correct recognition rate of classifiers can be directly employed as feature evaluation criteria. That is, besides the purpose of classification, inference engines are used for FS. This type of FS methods is categorized as the *wrapper* or the *embedded* one [1], [2], [12], [32]. A *wrapper* method [8], [32] trains certain classification model, and then evaluates a tested feature subset according to the performance of the trained model on the validation data. In an *embedded* model [3]–[6], [31], the parameter information of a trained or training classifier is used to evaluate the contribution of features. Generally, both the *wrapper* and the *embedded* methods can guarantee high classification accuracy, but they are usually computationally demanding [7]–[9]. These methods may also suffer from the problem of overfitting [7], [8]. Alternatively, many *filter* methods are developed in such a way that the FS process is independent of a classification learning procedure. In the *filter* methods, various criteria instead of classification error have been developed to measure the feature relevance, i.e., the relationship between features and class labels, and/or the feature dependence, i.e., the relationship among features. In [10], [11], the correlation or coefficient, the second-order statistics between two variables, was used to measure the feature relevance and/or the feature dependence. In [2], the concept of consistency was used to identify the salient features, behind which the rationale is that the data patterns close to each other are expected to have the same class label, and in [13]–[17], mutual information (MI) and probabilistic divergence were used for the purpose of FS. In [18], clustering information was exploited to find the appropriate features for classification. The studies on FS have been surveyed in many literatures [2], [19], [20]. It can be noted that most of the currently developed FS criteria of the *filter* models do not directly reflect the classification decision rule. Given that the major objective of FS is to raise the classification performance, it is desirable that the classification decision rule directly appears in the process of determining the useful features. This is the main motivation for proposing a new FS criterion in this paper. In order to reduce the misclassification risk, the classification decision rule is to assign a pattern (say, x) to the class having the maximum *a posteriori* probability, $p(c_i|x)$. In Bayesian decision theory, the Bayesian discriminant (BD) [21] can not only reflect this classification rule, but also can evaluate how likely that the decision, which is made according to this rule, is correct. Thus, in this paper, the concept of BD is explored as the core of the feature criterion. In this sense, the criterion is called BDFS for short. In BDFS, the *a posteriori* probabilities are firstly estimated by using Parzen window [22] and the Bayesian formula [21]. Then, the classification abilities of the features are evaluated by using these probability estimates.

Until now, heuristic strategies, such as backward and forward searching, are commonly used for searching features because they can be implemented in a relatively simple way. Also, the complexity of heuristic searching is $O(M^2)$, where M is the number of the given features. These heuristic strategies are clearly more computationally efficient compared with the exhaustive searching. However, with the increasing of M, the computational burden of these searching strategies becomes so heavy that it makes the application of these strategies difficult. Thus, researchers have been trying to make use of optimization strategies in the continuous space to conduct FS [18], [23], [24]. In this type of feature searching scheme, the weights of a FS criterion reflect the contributions of all the features. According to the optimization result of that selection criterion, the classification abilities of all the features can be evaluated simultaneously. Hitherto, many developed optimization approaches in the continuous space have been used for FS. For example, in [23], the similarity-based criterion was optimized through applying Lagrange multipliers. In [18], a type of least-square

Manuscript received March 15, 2004; revised September 3, 2004. This work was supported by the Hong Kong SAR Government CERG under Project 9040601-570. This paper was recommended by Associate Editor Y.-K. Chen.

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

Digital Object Identifier 10.1109/TCSI.2005.844364

approach was used to optimize a clustering-information-based FS criterion. These continuous-space searching strategies are generally highly efficient especially when one is handling a large feature set. Thus, a searching strategy of this type is considered in our study. We developed a gradient-based algorithm to maximize BDFS. The experimental results will show that this algorithm is able to solve a FS problem with great efficiency.

To sum up, the main advantages of our study are listed as follows.

1) The classification ability is directly evaluated according to the generic data analysis. This guarantees excellent FS results compared with other methods.
2) The feature searching is carried out through optimizing the FS criterion in a continuous space.

The proposed algorithm can be implemented in a very simple way and exhibits high convergence rate. More importantly, the huge computational reduction of the proposed algorithm does not sacrifice the effectiveness of FS. In order to evaluate the proposed FS scheme, an extensive comparative work with other related schemes in terms of effectiveness and efficiency are presented. These schemes include pulsewidth MI (PWMI) [14], MI-raw-FS [15], quadratic MI FS (QMIFS) [16], the $Q-\alpha$ FS scheme ($Q-\alpha$-FS) [18] and a support vector machines (SVM) based FS scheme, SVM recursive feature elimination (SVM RFE) in [3]. The comparative results can clearly show the merits of the proposed scheme.

This paper is organized as follows. Section II gives the background of our study, including the Bayesian decision theory and Parzen window probability estimator. Then, the proposed FS criterion and the approach for optimizing this criterion are detailed in Section III and Section IV, respectively. After that, the study results are presented and discussed in Section V. Finally, the conclusions are drawn.

II. Background

A. Bayesian Decision Theory

The Bayesian decision theory [21] is a versatile statistical method for performing pattern classification. Suppose that we have a general classification problem having Nx objects. x_i and $c_i(1 \le i \le Nx)$ denote a data pattern and the classification label of that pattern, respectively. This is a L-class classification problem, i.e., $c_i = \omega_l, l = 1, 2, \ldots, L$.

Based on the decision rule of the Bayesian theory, x_i is assigned to class ω_m, where $\omega_m = \arg\max_{\mathrm{k}}(p_\Lambda(\omega_k|x_i))$. Λ represents the classifier model for delivering *a posteriori* probabilities $p(\omega_k|x)$. The risk of making the above decision can be measured by using BD. In the case of a two-class problem ($L = 2$), the BD of Λ is defined as

$$D = \sum_{i=1}^{Nx} d(x_i, \Lambda) = \sum_{i=1}^{Nx} \left(p_\Lambda(\omega = c_i|x_i) - p_\Lambda(\omega \neq c_i|x_i)\right). \tag{1}$$

For a multiclass problem ($L > 2$), BD is defined in different ways [25], [26]. For the sake of convenience, we define the BD as

$$D = \sum_{i=1}^{Nx} d(x_i, \Lambda) = \sum_{i=1}^{Nx} \log \frac{p_\Lambda(\omega = c_i|x_i)}{p_\Lambda(\omega \neq c_i|x_i)} \tag{2}$$

which can be used to solve either two-class or multiclass problem.

In a classification training process, the task is to increase the BD (1) or (2) as much as possible through modifying classifier Λ. Being the preprocess of a classification, FS naturally aims to maximize the BD through searching different combinations of features.

B. Parzen Window Estimator

The Parzen window is a popular type of probability estimator [22]. Before explaining the Parzen window, we denote J_k as the pattern size of class ω_k. It is obvious that $Nx = J_1 + J_2 + \ldots + J_L$. We rewrite the L-group classification data X mentioned in the previous section as

$$\{\underbrace{x_1^{(1)}, x_2^{(1)}, \ldots, x_{J1}^{(1)}}_{\text{the pattern belonging to class } \omega_1}, \ldots, \underbrace{x_1^{(k)}, x_2^{(k)}, \ldots, x_{Jk}^{(k)}}_{\text{the pattern belonging to class } \omega_k}, \ldots, \underbrace{x_1^{(L)}, x_2^{(L)}, \ldots, x_{JL}^{(L)}}_{\text{the pattern belonging to class } \omega_L}\}$$

Using this data set, the *a priori* probabilities can be evaluated by

$$p(x) = \frac{1}{Nx}\sum_{i=1}^{Nx} p(x|x_i) = \frac{1}{Nx}\sum_{i=1}^{Nx} \kappa(x - x_i, \Sigma_i) \tag{3}$$

$$P(\omega_k) = \frac{J_k}{Nx} \tag{4}$$

where κ is the kernel function of the Parzen window. The conditional probability $p(x|\omega_k)$ is

$$\begin{aligned} p(x|\omega_k) &= \frac{1}{J_k}\sum_i p\left(x|x_i^{(k)}\right) \\ &= \frac{1}{J_k}\sum_{i=1}^{J_k} \kappa\left(x - x_i^{(k)}, \Sigma_i^{(k)}\right), \qquad k = 1, 2, \ldots, L. \end{aligned} \tag{5}$$

Using the Bayesian formula [21], we have

$$p(\omega_k|x) = \frac{p(x|\omega_k)p(\omega_k)}{p(x)} = \frac{\sum_{i=1}^{J_k} \kappa\left(x - x_i^{(k)}, \Sigma_i^{(k)}\right)}{\sum_{i=1}^{Nx} \kappa(x - x_i, \Sigma_i)}. \tag{6}$$

In this paper, a symmetric Gaussian function is chosen as the kernel function κ.[1] The general form of the M-dimensional Gaussian function is

$$\begin{aligned} &\kappa(x - x_0, \Sigma_0) \\ &= G(x - x_0, \Sigma_0) \\ &= \frac{1}{(2\pi)^{\frac{M}{2}}||\Sigma_0||^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - x_0)\Sigma_0^{-1}(x - x_0)^T\right) \end{aligned} \tag{7}$$

where Σ_0 is determined from distribution of a given data. As the given data is preprocessed to have a unit variance and a zero mean in each coordinate, we set Σ_0 with a scalar. For a data point $x_0 \in X$, Σ_0 is set to the half of the mean square Euclidean

[1]Using the Gaussian function as the kernel of the Parzen window probability estimator also has its limitation—the estimation accuracy may be (greatly) reduced when a given problem has nonconvex classes, although Parzen window estimator theoretically can approximate the real underlying probability as much as possible. We are grateful to the anonymous referee who pointed this out to us.

distances between x_0 and all the data point to X, i.e., $\Sigma_0 = (1/(Nx-1))\sum_{j=1}^{Nx}(d(x_j,x_0))^2$, where $(d(x_j,x_0))^2 = (x_j - x_0)(x_j - x_0)^T$.

III. Bayesian Discrimant Feature Selection Criterion

A. Definition of BDFS

Based on (2), the BDFS criterion of a feature set S is defined as

$$\mathrm{BDFS}(S) = \frac{1}{Nx}\sum_{i=1}^{Nx}\log\frac{p_S(\omega = c_i|x_i)}{p_S(\omega \neq c_i|x_i)} \tag{8}$$

where $p_S(\omega = c_i|x_i)$ is the probability of the pattern x_i belonging to the class c_i, and $p_S(\omega \neq c_i|x_i)$ is the probability of the pattern x_i not belonging to the class c_i. Intuitively, BDFS measures how likely the patterns in X can be correctly classified when the feature subset S is used. A large value of BDFS indicates the high classification ability of S. Negative values of BDFS(S) mean that the classes of most patterns cannot be correctly recognized with the feature subset S.

For x_i, the whole data set X can be partitioned into two complementary groups—X_i including all the patterns having the same class label with x_i and $\overline{X}_i$ consisting of all the patterns but ones in X_i, i.e., $\overline{X}_i = X - X_i$. Based on the probability estimate (6), we have

$$\begin{aligned}\mathrm{BDFS}(S) &= \frac{1}{Nx}\sum_{i=1}^{Nx}\log\frac{p_S(\omega = c_i|x_i)}{p_S(\omega \neq c_i|x_i)} \\ &= \frac{1}{Nx}\sum_{i=1}^{Nx}\left(\underbrace{\log\sum_{x_k\in X_i}G(x_k - x_i,\Sigma_i)}_{A}\right. \\ &\qquad \left.\underbrace{-\log\sum_{x_k\in\overline{X}}G(x_k - x_i,\Sigma_i)}_{B}\right). \end{aligned}\tag{9}$$

The physical meaning of BDFS is evident. It is known that $G(x_k - x_i, \Sigma_i)$ enumerates the similarity of x_i to x_k, or the "drawing force" of x_k to x_i. Thus, part A of (9) measures the "drawing force" to x_i from the interclass samples, whereas part B of (9) measures the outer class "drawing force" to x_i. BDFS of x_i evaluates the extent that the interclass "drawing force" to x_i is larger than the intraclass one.

B. Comparison Between the MI Based Criteria and BDFS

The MI-based feature evaluation criteria have been described in literatures [13]–[16]. To estimate an MI-based criterion, the knowledge on the underlying conditional probability ($p(x)$ and $p(x|c)$) are required. In [14]–[16], the Parzen window conditional probability estimators (3)–(5) are employed for this purpose. These MI-based criteria are similar to the proposed BDFS. Thus, we compared BDFS with these MI based criteria in this section.

Based on the commonly used definition of mutual information

$$I_s(X;C) = \sum_{k=1}^{L}\int_x p(\omega_k,x)\log\frac{p(\omega_k,x)}{P(\omega_k)p(x)}dx$$

PWMI (10) and MI-raw (11) were proposed in [14] and [15], respectively.

$$\mathrm{PWMI} = H(C) + \frac{1}{Nx}\sum_{i=1}^{Nx}\sum_{k=1}^{L}p(\omega_k|x_i)\log p(\omega_k|x_i). \tag{10}$$

$$\mathrm{MI-raw} = \frac{1}{Nx}\sum_{i=1}^{Nx}\log\frac{p(x_i,c_i)}{p(x_i)P(c_i)}. \tag{11}$$

Most recently, Chow and Huang [16] employed the QMI (3.5) to effectively measure the relevancy of features.

$$\begin{aligned}\mathrm{QMI} = &\log\left(\sum_{\omega}\int p(\omega,x)^2dx\right) + \log\left(\sum_{\omega}P(\omega)^2\right) \\ &\times\left(\int p(x)^2dx\right) - 2\log\left(\sum_{\omega}\int p(\omega,x)P(\omega)p(x)dx\right). \end{aligned}\tag{12}$$

It is noted that PWMI (10) measures the class information distribution around each data sample. For a sample, when the class information around it is distributed along one class direction, PWMI achieves the maximum. When the class information around x is distributed equally in all class directions, PWMI achieves the minimum. QMI (12), a modified Euclidean distance of two probabilities, measures the difference between probabilities $p(c,x)$ and $p(x)p(c)$ [17]. All these ideas are useful for evaluating the classification ability of feature subset. PWMI and QMI, however, do not directly reflect the classification rule. As for MI-raw (11), it can be rewritten as

$$\begin{aligned}\mathrm{MI-raw} &= \frac{1}{Nx}\sum_{i=1}^{Nx}\left(\log p(\omega = c_i|x_i) - \log P(c_i)\right) \\ &= \frac{1}{Nx}\sum_{i=1}^{Nx}\left(\log p(\omega = c_i|x_i)\right) \\ &\quad - \sum_{k=1}^{L}P(\omega_k)\log\left(P(\omega_k)\right). \end{aligned}\tag{13}$$

The second part of (13), the entropy of the class, is a constant for any feature subset. It, thus, has no effect on the result of FS. The first part is the determinant part of MI-raw, and thus is called MI-raw1 below. MI-raw1 is similar with BDFS in a sense of format, and was particularly compared with BDFS. As shown in Fig. 1, MI-raw1 clearly biases the patterns with $p(\omega|x) < 0.5$. With the decreasing of $p(\omega|x)$, the absolute value of MI-raw1 increases rapidly. As a result, the patterns with a small $p(\omega|x)$ may have a determinant effect on MI-raw1. We know that noise samples always have a very small $p(\omega|x)$. Thus, a small fraction of noise samples may greatly reduce the value of MI-raw1 of an important feature. In that scenario, MI-raw1 (i.e., MI-raw) may not be able to differentiate an important feature from the irrelevant ones. On the other hand, the proposed BDFS does not

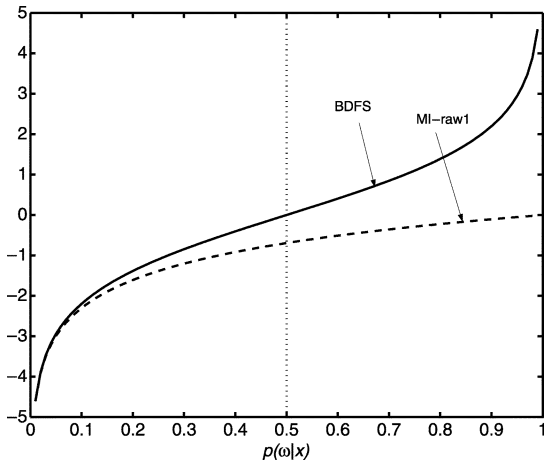


Fig. 1. Comparison between BDFS and MI-raw1- the first part of MI-raw (13).

suffer from this sort of problem because it balances the effects of the patterns with $p(\omega|x) < 0.5$ and ones with $p(\omega|x) > 0.5$. With this characteristic, BDFS is more robust than MI-raw. Below, we present synthetic examples to further elaborate this discussion.

A 10-dimension synthetic data set $\{X; C\}$ was firstly generated. The 100 data points in this data set evenly fall into two classes (i.e., $c = -1$ or 1). Among the 10 features, only x_1 and x_2 are relevant. For the first 50 points, x_1 was drawn from $cN(g, 1)$ and x_2 was drawn from $N(0, 1)$, while, for the other 50 points, x_1 and x_2 were drawn from $N(0, 1)$ and $cN(g, 1)$ respectively. As for the other eight features, all of them are irrelevant, and are randomly sampled from $N(0, 1)$. Under the forward searching scheme [13], [17], [19], BDFS and the MI based criteria, i.e., PWMI, MI-raw and QMI, are used to identify the two salient features. The correct selection order is $x_1 \rightarrow x_2$ or $x_2 \rightarrow x_1$. In each case, all the criteria were tested with 100 independently generated data sets. The presented results are the statistical results of 100 different trials.

We first tested different values of g in a range of [1.5, 3]. Obviously, with the decreasing of g, the overlapping between the two classes becomes larger, i.e., the number of noises increases. The selection accuracy rates are listed in Table I(a). In the case of the relative large overlapping, BDFS and QMI outperformed the other criteria. In a practical application, it is possible that the patterns are unintentionally mislabeled. To simulate this scenario, we set $g = 1$, then randomly selected certain points and changed the class labels of these points. Different numbers with "mislabeled" noises were tested. The correctness rates of 100 trials are listed in Table I(b). Obviously, these results show that BDFS and QMI are better than PWMI and MI-raw. In these simple cases the merits of BDFS have been shown. In Section V, more comparative results will be presented.

IV. Efficient Algorithm for Optimizing BDFS

In order to evaluate all the M features simultaneously, the M-elements vector $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_M]^T$ is introduced to modify BDFS, which satisfies

$$\alpha_i \in (0, 1) \text{ and } \sum_{i=1}^{M} \alpha_i = 1. \tag{14}$$

TABLE I
Comparison of Different FS Criteria on Synthetic Data Set. Values Listed in This Table are Correctness Rate of 100 Trials. (a) Results With Different δ^2. (b) Results With Different Size of Noises

(a)

g	1.5	1	0.8	0.5
PWMI	100%	99%	84%	51%
MI-raw	100%	95%	77%	39%
QMI	100%	96%	91%	55%
BDFS	100%	100%	94%	57%

(b)

The number of noise	0	10	15
PWMI	99%	70%	62%
MI-raw	95%	79%	62%
QMI	96%	87%	71%
BDFS	100%	86%	73%

The modified Gaussian function (7) and BDFS (8) are

$$G(x_k - x_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{M}{2}} \Sigma_i^{\frac{1}{2}}} \exp\left(-\frac{1}{2\Sigma_i} \sum_{j=1}^{M} \alpha_j (x_{kj} - x_{ij})^2\right) \tag{15}$$

$$\begin{aligned} \text{BDFS}(\alpha) &= \frac{1}{Nx} \sum_{i=1}^{Nx} \log \frac{\sum_{x_k \in X_i} G(x_k - x_i, \Sigma_i)}{\sum_{x_k \in \overline{X}_i} G(x_k - x_i, \Sigma_i)} \\ &= \frac{1}{Nx} \sum_{i=1}^{Nx} \log \frac{\sum_{x_k \in X_i} \exp\left(-\frac{1}{2\Sigma_i} \sum_{j=1}^{M} \alpha_j (x_{kj} - x_{ij})^2\right)}{\sum_{x_k \in \overline{X}_i} \exp\left(-\frac{1}{2\Sigma_i} \sum_{j=1}^{M} \alpha_j (x_{kj} - x_{ij})^2\right)}. \end{aligned} \tag{16}$$

In (16), the value of α_j reflects the importance of the feature f_j to BDFS. Assume that α_{opt} is the weight set that maximizes BDFS (16) under the constraint (15). A large element of α_{opt} indicates that the corresponding feature poses much effect on delivering a better classification result.

The gradient of BDFS (16) in terms of α_j is

$$\frac{\partial}{\partial \alpha_j} \text{BDFS} = \frac{1}{Nx} \sum_{i=1}^{Nx} (A_i - B_i) \tag{17}$$

where

$$A_i = \frac{\sum_{x_k \in \overline{X}_i} \left(\exp\left(-\frac{1}{2\Sigma_i} \sum_{j=1}^{M} \alpha_j (x_{kj} - x_{ij})^2\right) \frac{(x_{kj} - x_{ij})^2}{2\Sigma_i} \right)}{\sum_{x_k \in \overline{X}_i} \exp\left(-\frac{1}{2\Sigma_i} \sum_{j=1}^{M} \alpha_j (x_{kj} - x_{ij})^2\right)},$$

$$B_i = \frac{\sum_{x_k \in X_i} \left(\exp\left(-\frac{1}{2\Sigma_i} \sum_{j=1}^{M} \alpha_j (x_{kj} - x_{ij})^2\right) \frac{(x_{kj} - x_{ij})^2}{2\Sigma_i} \right)}{\sum_{x_k \in X_i} \exp\left(-\frac{1}{2\Sigma_i} \sum_{j=1}^{M} \alpha_j (x_{kj} - x_{ij})^2\right)}.$$

And our simple gradient based algorithm to fulfil the constraint optimization of BDFS can be stated as following.

Step 1) Set the values of $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_M]^T$ randomly, satisfying $\alpha_i \in (0,1)$ and $\sum_{i=1}^{M} \alpha_i = 1$.
Step 2) Calculate $\alpha'_j = \alpha_j + (\partial/\partial\alpha_j)\text{BDFS}$, $1 \le j \le M$, using (4.4).
Step 3) For any $\alpha'_k < 0$, set $\alpha'_k = 1/10M$.
Step 4) Obtain the new weight values through normalizing α'_j, i.e., $\alpha_j^{\text{new}} = \alpha'_j / \sum_{j=1}^{M} \alpha'_j$.
Step 5) Calculate the difference between α^{new} and $\alpha(\text{Diff})$, i.e., $\text{Diff} = \sum_{j=1}^{M} abs(a_j - a_j^{\text{new}})$.
Step 6) If Diff is small than 0.001 then go to Step 7
Else, $\alpha_j = \alpha_j^{\text{new}}$ for $1 \le j \le M$, then goto step 2. End If.
Step 7) Output α.

The computational complexity of this process is $O(HM)$, where M and H are the number of features and the number of cycle, respectively. Generally, this process stops within 10 cycles, which will be shown in the results of the next section. Thus, the proposed FS scheme is very efficient even when M is a large value.

V. Results and Discussion

In this section, BDFS-G represents the FS technique based on BDFS and the gradient based algorithm described in Section IV. We compared BDFS-G with other related techniques, such as, PWMI [14], MI-raw-FS [15], QMIFS [16], $Q-\alpha$-FS [18] and SVM-RFE [3]. As mentioned before, the first three models make use of Parzen window probability estimator, just like BDFS-G does. And similar to BDFS-G, $Q-\alpha$-FS employs the optimization approaches in the continuous space to weigh all the given features. SVM-RFE is a typical and efficient embedding FS model. Besides these methods, the BDFS based forward FS scheme (BDFS-F) was also implemented and compared with BDFS-G. Through these comparisons, the effectiveness and the efficiency of BDFS-G are extensively evaluated.

In a real classification example, in which no *a priori* knowledge is available, we evaluated the FS results by using the classification accuracy on the test data. Among several feature subsets of the same size, the one that delivers the highest classification accuracy is considered the best. Five classifiers were employed for this purpose. They are a neural network (NN) classifier, two SVM models, a decision tree (DT) and an k-NN rule. NN, DT and k-NN are available in Weka software (available at http://www.cs.waikato.ac.nz/~ml/weka), in which the parameters were set to the default values throughout our investigation. The employed SVM models are the SVM with "Linear" kernel (SVM-L) and the SVM with "RBF" kernel (SVM-R). These SVM models are available at http://www.isis.ecs.soton.ac.uk/resources/svminfo. All our studies were conducted by using Matlab 6.1 on Sun Ultra Enterprise 10 000 workstation having 100-MHz clock frequency and 4-GB memory.

Also, as the initialization may have an effect on the results of $Q-\alpha$-FS and BDFS-G, these schemes were tested with ten different trials in each example. The statistical results of the ten trials, i.e., the mean and the variance, are presented in this paper.

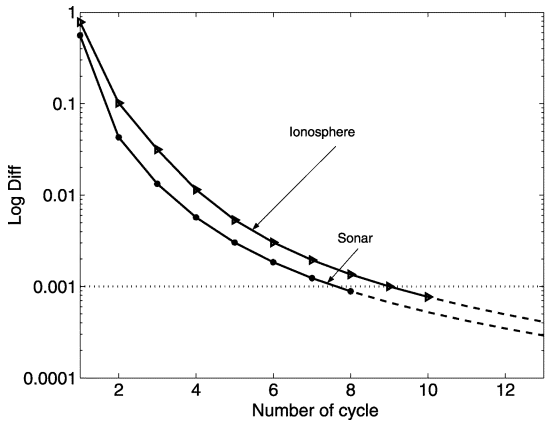


Fig. 2. Change trend of Diff in BDFS-G on the UCI classification data sets.

A. UCI Benchmarks

The FS methodologies were applied to the benchmarks of UCI data repository (available at http://www.ics.uci.edu/~mlearn/MLRepository.html). In these applications, all the forward FS schemes (including PWMI, QMIFS, MI-raw-FS, and BDFS-F) stopped when half of the original features had been selected.

Sonar Classification Data: This dataset was constructed to distinguish sonar returns bounced off a metal cylinder from those bounced off a rock. It consists of 104 training patterns and 104 testing patterns. It has 60 input features and two output classes, metal/rock.

Ionosphere Classification Data: In this binary classification task, the "bad" radar is to be detected on the basis of the collected data. This data have the 351 samples, each of which consists of 34 features. We randomly separated the whole data into two groups—200 for training and 151 for testing.

In these examples, BDFS-G exhibited very efficient performance because BDFS-G only required approximately 10 cycles to meet the stopping criterion, as shown in Fig. 2. In Fig. 3, the efficiency comparisons of different methods are illustrated. These comparative results indicate that BDFS-G and $Q-\alpha$-FS are much efficient than the others. This is primarily because both BDFS-G and $Q-\alpha$-FS employ the optimization approaches in the continuous space to conduct feature searching. Also, it is worth noting that, with the increasing of the number of sample, the computational complexity of SVM-RFE increases significantly faster than those of other methodologies.

In Tables II and III, different methodologies are compared in terms of FS effectiveness. It is firstly noted that, attributed to the advantages of BDFS, BDFS-G and BDFS-F can obtain better results than the others. BDFS-G consistently delivers much better performance compared with another efficient scheme such as $Q-\alpha$-FS. Also, the FS results of BDFS-G are comparable to or even better than those of BDFS-F. Thus, it can be concluded that, using the proposed gradient based feature searching strategy, the extent of the efficiency raised has no adverse effect on the effectiveness of FS. In Table IV, the results on other UCI datasets are briefed. These results similarly suggest that high performance can be delivered by the proposed BDFS-G. The comparative result between BDFS-G and BDFS-F show that BDFS-G has no particular advantage in terms of efficiency when the number of

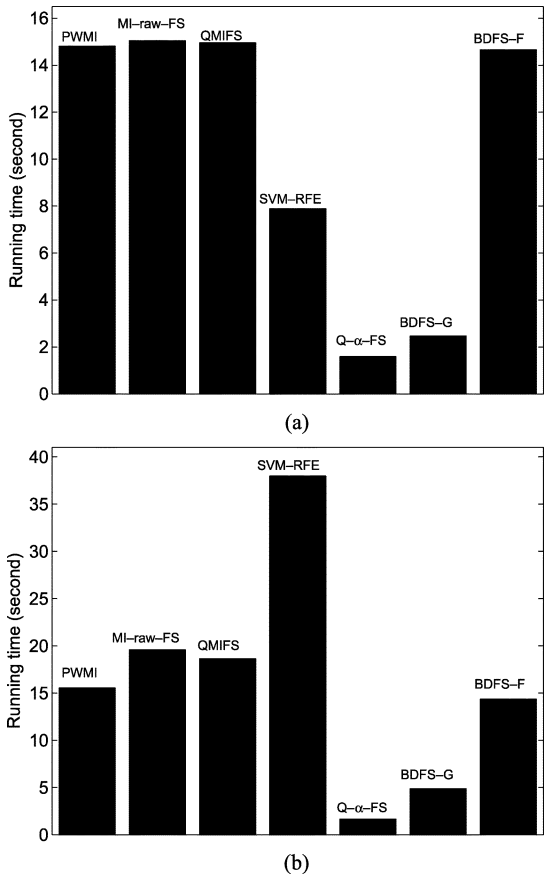


Fig. 3. Comparison of running time on UCI classification data sets. (a) Sonar data set. (b) Ionosphere data set.

TABLE II
COMPARISON OF SONAR CLASSIFICATION DATA SET. BEST RESULTS IN EACH CASE ARE HIGHLIGHTED IN BOLD FACE

Number of selected features		PWMI	MI-raw-FS	QMIFS	SVM -RFE	Q-α-FS	BDFS-F	BDFS-G
k-NN	4	0.78	0.67	0.76	0.76	0.60 ± 0	0.72	**0.78** ± 0
	8	0.83	0.77	**0.88**	0.85	0.78 ± 0.016	0.82	0.83 ± 0
	15	0.88	0.85	0.91	0.88	0.87 ± 0.008	0.89	**0.92** ± 0.012
	60	0.95						
SVM-R	4	0.78	0.68	0.70	0.82	0.59 ± 0	0.77	**0.82** ± 0
	8	0.84	0.77	**0.88**	0.81	0.73 ± 0.020	0.79	0.85 ± 0
	15	0.88	0.86	**0.90**	0.90	0.85 ± 0.025	0.88	0.88 ± 0.003
	60	0.93						
SVM-L	4	0.67	0.63	0.70	**0.74**	0.40 ± 0	0.69	0.71 ± 0
	8	0.69	0.71	0.69	0.75	0.64 ± 0.007	0.74	**0.75** ± 0
	15	0.74	0.69	0.74	0.73	0.73 ± 0.041	**0.77**	0.76 ± 0.006
	60	0.75						
NN	4	0.71	0.67	0.71	0.72	0.61 ± 0	0.74	**0.75** ± 0
	8	0.78	0.73	0.72	0.80	0.63 ± 0.023	**0.81**	0.76 ± 0
	15	0.79	0.79	0.83	0.77	0.84 ± 0.019	**0.84**	**0.84** ± 0.016
	60	0.87						
DT	4	**0.72**	0.63	0.70	0.65	0.40 ± 0	0.65	0.65 ± 0
	8	0.68	0.69	**0.73**	0.68	0.40 ± 0.015	0.71	0.65 ± 0
	15	0.75	0.76	0.76	0.71	0.75 ± 0.013	**0.76**	**0.76** ± 0.001
	60	0.73						

the selected features is less than 10. And the efficiency advantage of BDFS-G becomes significant with the increasing of the number of selected features. Generally, in real applications, it

TABLE III
COMPARISON OF IONOSPHERE DATA SET. BEST RESULTS IN EACH CASE ARE HIGHLIGHTED IN BOLD FACE

Number of selected features		PWMI	MI-raw-FS	QMIFS	Q-α-FS	SVM-RFE	BDFS-F	BDFS-G
k-NN	3	0.87	0.90	0.91	0.87 ± 0	0.85	0.90	**0.91** ± 0
	5	**0.89**	0.89	0.90	0.85 ± 0	0.89	0.86	0.87 ± 0
	9	0.89	0.88	0.88	0.85 ± 0	0.88	**0.88**	0.88 ± 0
	33	0.83						
SVM-R	3	0.90	0.90	0.92	0.84 ± 0	0.86	0.90	**0.92** ± 0
	5	0.89	0.84	**0.89**	0.86 ± 0	0.88	0.88	0.87 ± 0
	9	0.87	0.87	0.89	0.86 ± 0	0.90	0.88	**0.89** ± 0
	33	0.85						
SVM-L	3	0.81	0.81	**0.82**	0.71 ± 0	0.83	0.81	0.80 ± 0
	5	0.81	0.81	0.81	0.68 ± 0	0.87	**0.85**	**0.85** ± 0
	9	0.83	0.81	0.83	0.79 ± 0	0.87	0.82	**0.86** ± 0
	33	0.88						
NN	3	0.83	0.83	0.85	0.78 ± 0	0.89	0.89	**0.93** ± 0
	5	0.90	0.89	0.89	0.83 ± 0	0.92	0.89	**0.93** ± 0
	9	0.89	0.89	0.90	0.85 ± 0	0.91	0.88	**0.91** ± 0
	33	0.89						
DT	3	0.88	0.90	0.88	0.79 ± 0	0.85	**0.90**	**0.90** ± 0
	5	0.86	0.88	0.89	0.86 ± 0	0.88	**0.89**	**0.89** ± 0
	9	0.86	0.86	0.88	0.83 ± 0	0.88	**0.90**	0.88 ± 0
	33	0.87						

TABLE IV
COMPARISON OF OTHER UCI DATA SETS. (a) SUMMARY OF DATA SETS. (b) COMPARATIVE RESULTS. IN EACH COLUMN, UPPER VALUE IS CLASSIFICATION ACCURACY OF k-NN, AND LOW VALUE IS RUNNING TIME IN SECOND

(a)

Name of data	Number of feature	Number of class	Number of training patterns	Number of testing patterns
Vehicle	18	4	500	346
MUSK	166	2	200	6398
Isolet	617	26	300	1259

(b)

Data	Number of selected features	PWMI	MI-raw-FS	QMIFS	Q-α-FS	SVM-RFE	BDFS-F	BDFS-G
Vehicle	10	0.67 18	0.67 21	0.68 26	0.68 9	-	0.68 27	0.71 ± 0 22
MUSK	83	0.85 213	0.85 249	0.87 261	0.85 19	0.83 32	0.89 273	0.88 ± 0.02 19
Isolet	200	0.61 1.4×10^4	0.53 1.5×10^4	0.65 1.5×10^4	0.57 620	-	0.78 1.4×10^4	0.76 ± 0 1000

is required to select over 10 features. In that case, BDFS-G appeared to be the best option in terms of its effectiveness and efficiency.

We also evaluated the effect of initialization on the performance of BDFS-G. In practice, the close-zero standard deviations listed in Table II and Table III indicate that initialization has insignificant effect on the performance of BDFS-G. Further, for each feature, we compared the ordering results of 10 trails. In Table V, it summarizes the largest ordering difference and the top ranking of the feature(s) having this ordering difference in each example. For instance, in a sonar application, the largest ordering difference, i.e., 6, occurred in feature 60. The lowest rank of that feature is 31, and the highest one is 25. Based on the results presented in Table V, it can be concluded that the initialization has unnoticeable effect on the ordering results. Also, the less prominent features (i.e., those with relative lower ranking) are more likely to have a relatively large ranking change. Based on the above discussions, we can assert that the

TABLE V
SUMMARY OF BDFS-G ON UCI BENCHMARKS. FOR EACH EXAMPLE, RESULTS OF 10 TRIALS ARE COMPARED TO EACH OTHER. THESE COMPARATIVE RESULTS ARE USED TO VERIFY STABILITY OF BDFS-G

Name of data	The largest ranking difference	The top ranking of the feature associated with the largest difference
Sonar	6	25
Ionosphere	3	12
Glass	1	6
Vehicle	2	14
Isolet	4	311

results of BDFS-G are basically independent of initializations. This is very preferable for a gradient-based learning process.

B. Cancer Classification Using Microarray Gene Expression

In this study, the compared methods were applied to four different cancer classification microarray gene expression data. The forward FS schemes stopped after 100 features (genes) had been identified. SVM-RFE cannot deal with acute lymphoblastic leukemia (ALL) subtype classification data because this data includes more than twoclasses. The memory requirement makes it extremely difficult for $Q-\alpha$-FS to handle the data sets with over 10 000 features. Thus, $Q-\alpha$-FS was not applied to the examples of prostate cancer classification and ALL subtype classification.

Colon Tumor Classification: These data contain 62 samples collected from colon-cancer patients [27]. Among these samples, 40 samples are tumor, and 22 are labeled "normal." There are 2,000 genes selected based on the confidence in the measured expression levels. The whole sample set was randomly split into two disjoint parts—30 samples for the training and 32 for testing.

ALL-AML Classification: This classification task is aimed to distinguish the two variants of leukemia, namely, ALL and acute myelod leukemia (AML) [28]. The given training and test data sets consist of 38 and 34 samples, respectively. The expression levels of the 7129 genes are given in each sample.

Prostate Cancer Classification: The objective of this classification is to distinguish prostate cancer cases from noncancer cases [29]. This data set is available on http://www-genome.wi.mit.edu/mpr/prostate. There are 136 data samples, and each sample contains 12 600 features. We randomly partitioned the whole data set into two subsets. One 76 patterns was used for training, and the other 60 patterns were used for testing.

Subtype of ALL Classification Data Set: The task of this application is to correctly diagnosis the subtypes of the pediatric ALL [30], which is crucial because different subtypes have different treatment plan. The original data has been divided into six diagnostic groups (BCR-ABL, E2A-PBX1, Hyperdiploid>50, MLL, T-ALL and TEL-AML1), and a miscellaneous class that contains diagnostic samples that did not fit into any one of the above groups (thus labeled as "Others"). There are a total of 12 558 features and 327 samples in this dataset. This dataset has been partitioned into two disjoint subsets, of which 215 samples were used for training and 112 were used for testing.

The typical Diff trends of BDFS-G in these applications are illustrated in Fig. 4. And the comparisons on running time are

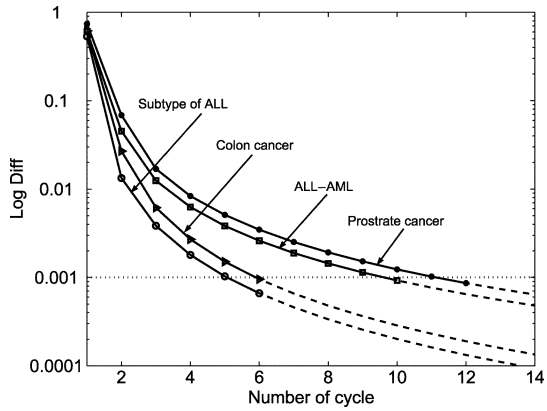


Fig. 4. Change of Diff in BDFS-G on the four microarray based cancer classification data sets.

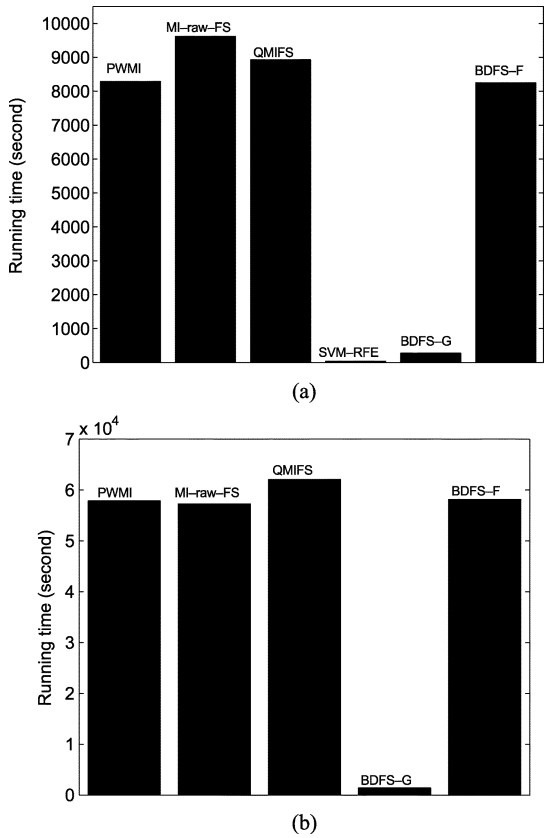


Fig. 5. Comparison of running time on microarray based cancer classification data sets. (a) Prostate; (b) Subtype of ALL.

presented in Fig. 5. All these results show that the proposed BDFS-G exhibits the quite fast convergence rate. For example, in the prostate cancer classification data, BDFS-G required 276 s to rank all the 12 000 features, while it took approximately 8245 s for BDFS-F to determine the top 100 important features, as presented in Fig. 5(a). In the ALL subtype classification example, the results illustrated in Fig. 5(b) show that the computational complexity of BDFS-F is more than 40 times of that of BDFS-G. Apparently, the high efficiency of BDFS-G is an appealing advantage when a huge feature set is given.

TABLE VI
COMPARISON OF COLON CANCER CLASSIFICATION DATA SET. BEST RESULT IN EACH CASE IS HIGHLIGHTED IN BOLD FACE

Number of selected features		PWMI	MI-raw-FS	QMIFS	Q-α-FS	SVM-RFE	BDFS-F	BDFS-G
k-NN	2	0.72	0.69	0.69	0.63 ± 0	**0.84**	0.78	0.81 ± 0
	4	0.78	0.69	0.78	0.50 ± 0	0.72	**0.91**	**0.91** ± 0
	8	0.81	0.81	0.84	0.44 ± 0	0.72	**0.94**	0.81 ± 0
SVM-R	2	0.81	0.72	0.75	0.63 ± 0	0.91	0.84	**0.88** ± 0
	4	0.84	0.75	0.84	0.59 ± 0	0.75	**0.91**	**0.91** ± 0
	8	0.63	0.75	0.88	0.59 ± 0	0.72	**0.94**	0.81 ± 0
SVM-L	2	0.84	0.88	0.78	0.47 ± 0	**0.91**	0.84	0.72 ± 0
	4	0.81	0.78	0.81	0.69 ± 0	0.81	**0.88**	**0.88** ± 0
	8	0.81	0.88	0.75	0.69 ± 0	0.81	**0.88**	**0.88** ± 0
NN	2	0.88	0.72	0.78	0.47 ± 0	0.88	**0.88**	0.75 ± 0
	4	0.81	0.75	**0.90**	0.50 ± 0	0.81	0.88	0.75 ± 0
	8	0.75	0.75	**0.88**	0.50 ± 0	0.84	0.84	0.84 ± 0
DT	2	0.72	0.75	0.72	0.69 ± 0	0.78	**0.75**	0.50 ± 0
	4	0.75	0.75	0.78	0.69 ± 0	0.75	0.75	**0.91** ± 0
	8	0.72	0.75	0.78	0.69 ± 0	0.75	0.72	**0.91** ± 0

TABLE VII
COMPARISON OF ALL-AML CLASSIFICATION DATA SET. BEST RESULT IN EACH CASE IS HIGHLIGHTED IN BOLD FACE

Number of selected features		PWMI	MI-raw-FS	QMIFS	Q-α-FS	SVM-RFE	BDFS-F	BDFS-G
k-NN	2	0.94	**0.94**	0.79	0.71 ± 0	0.91	0.76	0.91 ± 0
	4	0.91	0.91	0.91	0.79 ± 0	0.79	0.82	**0.94** ± 0
	8	0.82	0.82	0.97	0.76 ± 0	0.82	0.91	**0.97** ± 0
SVM-R	2	0.91	**0.91**	0.83	0.62 ± 0	0.88	0.79	0.88± 0
	4	0.88	0.88	0.88	0.65 ± 0	0.79	0.82	**0.94** ± 0
	8	0.79	0.79	0.91	0.62 ± 0	0.82	0.88	**0.97** ± 0
SVM-L	2	0.88	0.88	0.76	0.50 ± 0	0.88	0.79	**0.94** ± 0
	4	0.91	0.91	**0.91**	0.62 ± 0	0.85	0.76	0.76 ± 0
	8	0.59	0.59	0.59	0.62 ± 0	**0.88**	0.79	0.79 ± 0
NN	2	0.76	0.76	0.79	0.62 ± 0	0.88	0.79	**0.94** ± 0
	4	0.82	0.82	0.88	0.62 ± 0	0.79	0.76	**0.91** ± 0
	8	0.88	0.88	0.94	0.62 ± 0	0.79	0.85	**0.97** ± 0
DT	2	0.73	0.73	0.79	0.59 ± 0	0.88	**0.88**	0.82 ± 0
	4	0.76	0.73	0.79	0.59 ± 0	0.88	0.91	**0.97** ± 0
	8	0.91	0.91	0.79	0.62 ± 0	0.88	0.94	**0.97** ± 0

TABLE VIII
COMPARISON OF PROSTATE CANCER CLASSIFICATION DATA SET. BEST RESULT IN EACH CASE IS HIGHLIGHTED IN BOLD FACE

Number of selected features		PWMI	MI-raw-FS	QMIFS	SVM-RFE	BDFS-F	BDFS-G
k-NN	4	0.80	0.77	0.78	0.82	**0.85**	0.78 ± 0
	8	0.82	0.72	**0.85**	0.78	0.83	0.80 ± 0
	16	0.80	0.80	0.82	0.83	**0.92**	0.85 ± 0
SVM-R	4	0.83	0.80	0.82	0.82	**0.83**	0.68 ± 0
	8	0.82	0.77	0.85	0.82	**0.85**	**0.85** ± 0
	16	0.82	0.83	0.82	0.80	**0.90**	0.85 ± 0
SVM-L	4	0.73	0.72	0.72	0.83	**0.87**	0.65 ± 0
	8	0.68	0.73	0.73	0.75	**0.92**	0.72 ± 0
	16	0.72	0.73	0.73	0.77	**0.78**	0.77 ± 0
NN	4	0.85	0.75	**0.90**	0.83	0.87	0.75 ± 0
	8	0.78	0.72	0.80	0.75	**0.92**	0.78 ± 0
	16	0.80	0.85	0.77	0.82	**0.88**	0.83 ± 0
DT	4	0.77	0.77	0.83	0.80	**0.87**	0.80 ± 0
	8	0.80	0.77	0.85	0.78	**0.92**	0.82 ± 0
	16	0.75	0.75	0.80	0.67	**0.87**	0.87 ± 0

TABLE IX
COMPARISON OF SUBTYPES OF ALL CLASSIFICATION DATA SET. BEST RESULT IN EACH CASE IS HIGHLIGHTED IN BOLD FACE

Number of selected features		PWMI	MI-raw-FS	QMIFS	BDFS-F	BDFS-G
k-NN	8	0.79	0.77	0.78	**0.80**	0.79 ± 0
	16	0.81	0.79	0.77	**0.88**	0.81± 0
	24	0.83	0.77	0.83	0.88	**0.89** ± 0
NN	8	0.70	**0.85**	0.71	0.84	0.71 ± 0
	16	**0.83**	0.79	0.79	0.67	0.80 ± 0
	24	0.83	0.84	0.86	**0.90**	0.89 ± 0
DT	8	**0.76**	0.63	0.71	0.63	0.70 ± 0
	16	0.76	0.63	0.77	**0.83**	0.77 ± 0
	24	0.73	0.65	0.77	**0.83**	0.79 ± 0

TABLE X
SUMMARY OF BDFS-G RESULTS ON FOUR CANCER CLASSIFICATION DATA SETS. ONLY THE FIRST 50 FEATURE SUBSETS DETERMINED BY BDFS-G ARE CONSIDERED DURING SUMMARIZING

Name of data	Type of classifier	The highest classification accuracy	The smallest size of feature subset corresponding to the best classification result
Colon cancer	k-NN	0.91 ± 0	4
	SVM-R	0.91 ± 0	4
	SVM-L	0.88 ± 0	3
	NN	0.88 ± 0	4
	DT	0.91 ± 0	4
ALL-AML	k-NN	0.97 ± 0	6
	SVM-R	0.97 ± 0	5
	SVM-L	0.97 ± 0	3
	NN	0.88 ± 0	4
	DT	0.91 ± 0	4
Prostate cancer	k-NN	0.87 ± 0	18
	SVM-R	0.90 ± 0	36
	SVM-L	0.90 ± 0	14
	NN	0.87 ± 0	19
	DT	0.87 ± 0	12
Subtype of ALL	k-NN	0.94 ± 0	38
	NN	0.82 ± 0	30
	DT	0.81 ± 0	26

As for evaluating the effectiveness of these methods, comparative results are presented in Tables VI–IX. Also, the best classification results are summarized in Table X. Based on these results, we can draw similar conclusions with the cases of the UCI data. Firstly, in terms of FS effectiveness, both BDFS-G and BDFS-F outperformed other methods in most scenarios. Secondly, zero standard deviations of BDFS-G listed in all these tables indicate the stable performance of BDFS-G. Thus, we can assert that the BDFS-G is able to consistently deliver excellent results irrespective of the size of the data set.

VI. CONCLUSION

In this paper, a new type of *filter* model for performing FS is presented. Two important issues are addressed in an effort to enhance the FS performance. First, a new FS criterion is introduced. By using the concept of BD, this criterion can evaluate the classification ability of a feature subset in a straightforward way. This makes the proposed criterion very effective compared with other FS criteria. Second, we use a gradient based searching scheme to tackle the FS problem. This searching

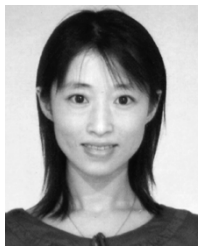
scheme is remarkably efficient even when it is used to handle a huge feature set. More importantly, the convergence of this gradient-based process is very robust to initialization. In this study, extensive examples are used to evaluate the proposed criterion and the searching scheme. The presented results clearly show the advantages of the proposed method.

ACKNOWLEDGMENT

The authors wish to express their sincere thanks to Dr. L. Wolf, Hebrew University for providing the source code of the $Q - \alpha$ method; and to the anonymous referees for providing useful comments to this paper.

REFERENCES

[1] J. W. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Mateo, CA: Morgan Kaufmann, 2001.
[2] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. London, U.K.: Kluwer , 1998.
[3] I. Guyon, J. Weston, and S. Barnhill, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 2, pp. 389–422, 2002.
[4] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in *Proc. NIPS*, 2000, pp. 668–671.
[5] R. Setiono and H. Liu, "Neural-network feature selector," *IEEE Trans. Neural Netw.*, vol. 8, no. 3, pp. 654–661, May 1997.
[6] D. MacKay, "A practical Bayesian framework for backpropagation networks," *Neur. Comput.*, vol. 4, pp. 448–472, 1992.
[7] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learning Res.*, vol. 3, pp. 1157–1183, 2003.
[8] E. P. Xing, M. I. Jordan, and R. M. Karp, "Feature selection for high-dimensional genomic microarray data," in *Proc. 18th Int. Conf. Machine Learning*, 2001, pp. 601–608.
[9] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Proc. 11th Int. Conf.*, San Francisco, CA, 1994.
[10] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. thesis, Dept. Comp. Sci., Waikato Univ., Waikato, New Zealand, 1999.
[11] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 301–312, Mar. 2002.
[12] A. L. Blum and P. Langley, "Selection of relevant feature and examples in machine learning," *Artif. Intell.*, vol. 97, no. 1–2, pp. 245–271, 1997.
[13] 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.
[14] 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.
[15] B. Bonnlander, "Nonparametric Selection of Input Variables for Connectionist Learning," Ph.D. thesis, Dept. Comp. Sci., Univ. of Colorado , Boulder, 1996.
[16] T. W. S. Chow and D. Huang, "Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information," *IEEE Trans. Neural Netw.*, vol. 16, no. 1, pp. 213–224, Jan. 2005.
[17] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1982.
[18] L. Wolf and A. Shashua, "Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weighted-based approach," Hebrew Univ., Jerusalem, Israel, Tech. Report 2003-58, , Jun. 2003.
[19] Feature selection algorithms: A survey and experimental evaluation, L. C. Molina, L. Belanche, and A. Nebot. (2002). [Online]. Available: http://www.lsi.upc.es/dept/techreps html/R02-62.html
[20] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers," *Patt. Recogn.*, vol. 33, pp. 25–41, 2000.
[21] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: Wiley, 2001.
[22] E. Parzen, "On the estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, pp. 1064–1076, 1962.
[23] B. Lazzerini and F. Marcelloni, "Feature selection based on similarity," *Electron. Lett.*, vol. 38, no. 3, pp. 121–122, Jan. 2001.
[24] S. K. Pal, R. K. De, and J. Basak, "Unsupervised feature evaluation: a neuro-fuzzy approach," *IEEE. Trans. Neur. Netw.*, vol. 11, no. 2, pp. 366–376, Mar. 2000.
[25] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, vol. 40, no. 12, pp. 3043–3054, Dec. 1992.
[26] S. Amari, "A theory of adaptive pattern classifiers," *IEEE Trans. Electron. Comp.*, vol. EC-16, no. 3, pp. 299–307, Mar. 1967.
[27] U. Alon, N. Barkar, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci. USA. Cell Biol*, vol. 96, no. 12, pp. 6745–6750, 1999.
[28] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, 1999.
[29] L. J. Veer, H. Dai, M. J. Vijver, Y. He, A. A. Hart, M. Mao, H. L. Peterse, K. V. Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530–536, 2002.
[30] E. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C. Pui, W. E. Evans, C. Naeve, L. Wong, and J. R. Downing, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, pp. 133–143, 2002.
[31] D. W. Aha and R. L. Bankert, "A comparative evaluation of sequential feature selection algorithms," in *Proc. 5th Int. Workshop on Artificial Intelligence and Statistics*, Ft. Lauderdale, FL, 1995, pp. 1–7.
[32] R. Kohavi and G. H. John, "The wrapper approach," in *Feature Extraction, Construction and Selection*, H. Liu and H. Motoda, Eds. Norwell, MA: Kluwer , 1998, pp. 33–50.



D. Huang received the B.E. degree from the Liaoning Institute of Technology, Jinzhou, China, and the M.E. degree from the Shenyang Institute of Automatics, Chinese Academy of Science, Shenyang, China, in 1993 and 1996, respectively. She is now working toward the Ph.D degree in electronic engineering at the City University of Hong Kong, Hong Kong, China.

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

Tommy W. S. Chow (M'93–SM'03) received the B.Sc. (First Hons.) and Ph.D. degrees from the University of Sunderland, Sunderland, U.K., in 1984 and 1988, respectively.

For his doctoral study, he worked on a collaborative project between The International Research and Development, Newcastle Upon Tyne, U.K, and the Ministry of Defense, (Navy), U.K. He joined the City University of Hong Kong, Hong Kong, as a Lecturer in 1988, where he is currently a Professor in the Department of Electronic Engineering. He has been a Consultant to the Mass Transit Railway, Kowloon-Canton Railway Corporation, Hong Kong. He has conducted other collaborative projects with the Kong Electric Co. Ltd, and Royal Observatory Hong Kong, and the MTR Hong Kong on the application of neural networks for machine fault detection and forecasting. Currently, he is collaborating with the Laboratories of Information Technology, National Singapore University on Bioinformatics. His main research has been in the area of neural network, learning theory, system identification, and machine fault diagnostics. He is an author and coauthor of numerous published works, including book chapters, and over 90 Journal articles related to his research.

Prof. Chow's research led to the Best Paper Award in the 2002 IEEE Industrial Electronics Society Annual meeting, Seville, Spain. He was the Chairman of Hong Kong Institute of Engineers, Control Automation and Instrumentation Division 1997–1998.