Discriminative Feature Selection by Nonparametric Bayes Error Minimization

Shuang-Hong Yang and Bao-Gang Hu, *Senior Member*, *IEEE*

Abstract—Feature selection is fundamental to knowledge discovery from massive amount of high-dimensional data. In an effort to establish theoretical justification for feature selection algorithms, this paper presents a theoretically optimal criterion, namely, the *discriminative optimal criterion* (DoC) for feature selection. Compared with the existing *representative optimal criterion* (RoC, [21]) which retains maximum information for modeling the relationship between input and output variables, DoC is pragmatically advantageous because it attempts to directly maximize the classification accuracy and naturally reflects the Bayes error in the objective. To make DoC computationally tractable for practical tasks, we propose an algorithmic framework, which selects a subset of features by minimizing the Bayes error rate estimated by a nonparametric estimator. A set of existing algorithms as well as new ones can be derived naturally from this framework. As an example, we show that the Relief algorithm [20] greedily attempts to minimize the Bayes error estimated by the k-Nearest-Neighbor (kNN) method. This new interpretation insightfully reveals the secret behind the family of *margin*-based feature selection algorithms [28], [14] and also offers a principled way to establish new alternatives for performance improvement. In particular, by exploiting the proposed framework, we establish the Parzen-Relief (P-Relief) algorithm based on Parzen window estimator, and the MAP-Relief (M-Relief) which integrates label distribution into the max-margin objective to effectively handle imbalanced and multiclass data. Experiments on various benchmark data sets demonstrate the effectiveness of the proposed algorithms.

Index Terms—Feature selection, discriminative optimal criterion, feature weighting.

1 Introduction

Feature subset selection is a preprocessing process aimed at identifying a small subset of highly predictive features out of a large set of raw input variables that are possibly irrelevant or redundant [15], [12]. It plays a fundamental role in the success of many learning tasks where high dimensionality arises as a big challenge, most evidently, in pattern recognition [27], [7], knowledge discovery and data mining [16], [42], information retrieval [9], [43], computer vision [4], [1], bioinformatics [34], [6] and so forth. The effects of feature selection [16], [15] have been widely recognized for its abilities in, e.g., facilitating data interpretation, reducing measurement and storage requirements, increasing processing speeds, defying curse of dimensionality and improving generalization performance, etc.

In this paper, feature selection is investigated as it is applied to classification scenarios. Suppose we are given a set of input vectors $\{\mathbf{x}_n\}_{n=1}^N$ along with corresponding targets (labels) $\{y_n\}_{n=1}^N$ drawn $i.i.d$ from an unknown distribution $p(\mathbf{x}, y)$, where $\mathbf{x}_n \in \mathbf{X} \subset \mathcal{R}^D$ is a training instance and $y_n \in \mathbf{Y} = \{0, 1, \dots, C-1\}$ is its label, N, D, C denote the training set size, the input space dimensionality and the total number of categories, respectively. The dth feature of $\mathbf{x}$ is denoted as $x^{(d)}$, $d = 1, 2, \dots, D$. The goal of feature selection is to select a subset of M ($M \ll D$) most predictive features, i.e., to find a preprocessing of data $\tau(\mathbf{x}) : \mathbf{x} \to (\mathbf{x} * \tau)$, where $\tau = [\tau_1, \dots, \tau_D] \in \mathcal{S} = \{0, 1\}^D$ with $||\tau||_0 = M$ is an operator selecting up to M elements from a D-vector, $(\mathbf{x} * \tau) = [x^{(1)}\tau_1, \dots, x^{(D)}\tau_D]$ denotes the element-wise product and $|| \cdot ||_0$ the L_0 norm. Let the feature selection criterion function be represented by $J(\cdot)$. Formally, the problem of feature selection can be formalized as the following optimization task:

$$\tau = \arg \max_{\tau \in \mathcal{S}, ||\tau||_0 = M} J(\tau). \tag{1}$$

From a methodological perspective, existing feature selection algorithms can be generally divided into two categories [15], [16]: *Wrapper* and *Filter*. The *Wrapper* (or *Embedded*) methods use a specific type of classifier to assess the quality of a feature subset, and select the optimal feature subset by minimizing the training error of the chosen classifier. In contrast, the *Filter* methods evaluate features based on certain criteria that are independent of any classifier and find the features by optimizing such criteria. In practice, while the Wrapper methods have been frequently observed to achieve better classification accuracy for the classifier being choosed, they are also much more time-consuming; and due to their dependence to the targeted classifier, they usually provide less generic knowledge of the data.

A feature selection method typically consists of two basic elements [10], [19]:

- *Evaluation criterion*, a measure to assess the goodness of a feature subset τ.

- *S.-H. Yang is with the College of Computing, Georgia Institute of Technology, 266 Ferst Drive, Atlanta, GA 30318. E-mail: shy@gatech.edu.*
- *B.-G. Hu is with the National Laboratory of Pattern Recognition (NLPR), Chinese Academy of Sciences, 95 ZhongGuanCun East Road, Beijing 100190, China. E-mail: hubg@nlpr.ia.ac.cn.*

Manuscript received 23 Feb. 2010; revised 15 Dec. 2010; accepted 15 Mar. 2011; published online 4 Apr. 2011.
Recommended for acceptance by V. Kumar.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2010-02-0106.
Digital Object Identifier no. 10.1109/TKDE.2011.72.

1041-4347/12/$31.00 © 2012 IEEE Published by the IEEE Computer Society

- *Search strategy*, a procedure to generate candidate subset τ for searching the feature power set.

Over the past decades, a large number of algorithms have been proposed for feature selection. However, most of the existing evaluation criteria are based on heuristic intuitions or domain knowledge. For example, the *Relief* algorithm [20] is recently interpreted as a method that optimizes the average heuristical margin [28], [14], [5], although the secret behind the definition of the margin is unclear. To deepen our understanding of existing feature selection methods and lay guidelines for the development of new algorithms, it is highly desirable to establish feature selection objectives (i.e., evaluation criteria) that possess sound theoretic understanding. To the best of our knowledge, the first effort for the theoretical treatment of feature selection is the so-called "Optimal Feature Selection" framework [21]. This work not only laid a solid foundation for feature selection, but also inspired the establishment of many successful algorithms [34], [44], [24]. However, the framework in [21] is *representative*, i.e., it attempts to minimize the information loss for *modeling* the relationship (e.g., the posterior distribution) between input and output in the process of dimensionality reduction. In practice, preserving the representative information to small details could be wasteful of training resources as our ultimate goal is merely classifying the data coarsely to several categories. Also, this representative framework could be practically risky as modeling posteriors is a nontrivial task especially when the training data is rare. Fortunately, we show in this paper that the "representative" framework is not the only option, and establish a "discriminative" framework for learning feature selectors. Unlike the representative framework, a *discriminative* counterpart considers classification directly by optimizing the ability of the feature selector in *discriminating* data from different classes. The discriminative framework is more advantageous in practice because by directly optimizing the discriminativeness (which relates to the difference of posteriors), it avoids direct modeling of the posteriors.

Given a certain evaluation criterion, feature selection is reduced to a combinatorial search problem, where each state in the search space is a possible feature subset. To achieve the optimal subset is NP-hard since the search space is exponential in the number of features. Therefore, heuristic search procedures are necessary. Existing search strategies are generally divided [19] into three classes: *complete* (exhaustive, best first, branch and bound, beam, etc.), *heuristic* (forward/backward sequential, greedy, etc.), or *stochastic* (simulated annealing, genetic algorithm, etc.). The recently proposed feature weighting procedure [33], [5], [43] is a greedy search strategy. It assigns to each feature a real valued number to indicate its usefulness, making it possible to efficiently search the feature power set simply by searching in a continuous space. For this reason, this paper will fix the search strategy at feature weighting and focus mainly on the aspect of evaluation criterion. However, extensions to other search schemes are straightforward (e.g., [38]).

In this paper, we first present a theoretically optimal criterion for feature selection, namely the *discriminative optimal criterion (DoC)*, as a complementarity to the representative one (referred to as *representative optimal criterion (RoC)* [21]). The DoC directly attempts to maximize the classification accuracy and naturally reflects the Bayes error in the objective. Compared to RoC, DoC is practically favorable if our ultimate goal is for the purpose of supervised classification. However, DoC itself is computationally intractable as it involves unknown probabilistic densities. To make DoC practical, we then propose an algorithmic framework for feature selection, which selects a subset of features by minimizing the Bayes error estimated by a nonparametric estimator. Many existing approaches as well as new ones [38], [39] can be naturally derived from this framework. For example, taking feature weighting as an example search strategy, we show that the Relief algorithm attempts to greedily minimize the nonparametric Bayes error that is estimated by k-nearest-neighbor (kNN) method. This new interpretation of Relief not only reveals the secret behind the margin definition in margin-based feature selection algorithms [28], [14], [5], but also enables us to identify the weaknesses of these methods so as to establish new algorithms to mitigate the drawbacks. In this paper, an alternative algorithm, called Parzen-Relief (P-Relief), is proposed, which resembles the standard Relief algorithm, but instead of kNN, it uses the Parzen window method to estimate the Bayes error. We show that the empirical performance of Parzen-Relief usually outperforms Relief. In addition, we find that the Relief makes an implicit assumption that the class distribution is balanced among every "one-versus-rest" split of the data. This undesirable assumption rarely holds in practice and substantially limits the performance of Relief in handling imbalanced or multiclass data set. To mitigate this drawback, we derive a MAP-Relief (M-Relief) algorithm based on the proposed algorithmic framework, which incorporates the class distribution into the margin maximization objective function and thus effectively captures the imbalanceness among classes. Both Parzen-Relief and MAP-Relief are of the same computational complexity as the standard Relief algorithm; yet, both of them demonstrate significant performance improvement over Relief as illustrated by our experiments on various benchmark data sets.

The organization of this paper is as follows: we briefly review the related work in Section 2 and then present the discriminative framework for feature selection in Section 3. Section 4 offers a new interpretation for Relief, and furthermore establishes two alternatives, i.e., the Parzen-Relief and MAP-Relief algorithm. In Section 5, we present the experiments and empirical results on both UCI data and large-scale real-world tasks. Finally, in Section 6, we summarize the whole paper.

2 Related Work

2.1 Theoretically Optimal Feature Selection

The "optimal feature selection" framework [21], for the first time, places a sound theoretical foundation for the feature selection task. Based on the information theory, this framework defines the optimality of a feature subset in the sense that it retains the most amount of information required for modeling the dependence between the input variables (features) and output variable (label) in the reduced-dimensional space. Let $\tau(\mathbf{x})$ denote the representation of $\mathbf{x}$ after the dimensionality reduction defined by τ, this framework requires that the posterior $p(y|\tau(\mathbf{x}))$ be as close as possible to the original one $p(y|\mathbf{x})$:

$$\begin{aligned}&\min_{\tau} KL\{p(y|\mathbf{x})||p(y|\tau(\mathbf{x}))\}\\&s.t.: ||\tau||_0 = M,\end{aligned} \tag{2}$$

where $KL\{p(\mathbf{x})||q(\mathbf{x})\} = E_X[p(\mathbf{x})\log\frac{p(\mathbf{x})}{q(\mathbf{x})}]$ denotes the Kullback-Leibler (KL) divergence between two distribution $p(\mathbf{x})$ and $q(\mathbf{x})$. We refer to this criterion as *representative optimal criterion* to emphasize its goal of retaining the information for modeling the relationship between the input and output. A variety of existing criteria could be seen as instances of RoC. For example, the entropy or mutual-information (MI) criterion and its variations [15], and also the objectives used in [29], [44], [34], [24], [43], etc. However, in the context of supervised classification, RoC might unnecessarily complicate matters because modeling posterior distributions for high-dimensional data is nontrivially more challenging than classifying the data coarsely to several classes. Clearly, if our ultimate goal is for classification purpose, an obvious objective is the classification accuracy; and there is no reason for going beyond this objective for more complicated ones.

2.2 Existing Discriminative Methods

There have been attempts for discriminative methodology of feature selection. Most of them are focused on developing algorithms by exploiting certain heuristic perspectives or domain knowledge, and hence significantly different from what we present in this paper as our emphasis here is to establish a general framework for discriminative feature selection from a theoretically optimal perspective.

In [27], Saon and Padmanabhan proposed to reduce the input dimensionality by approximately minimizing the Bayes error. They established algorithms for constructing linear feature transformation by maximizing the *average pairwise divergence* or minimizing the *union Bhattacharyya error bounds* and proved that these optimizations asymptotically relates to Bayes error minimization. To make their approach tractable, Gaussian assumption has to be made for the class-conditional densities, which, however, limits the performance of their methods for multimode non-Gaussian data that are quite common in practical applications. This limitation is also shared by similar methods such as the Bhattacharyya distance approach [7], [35].

In the context of vision recognition, Vasconcelos [31] proposed a feature selection approach based on the infomax principle. Although their criterion was tied to the Bayes error, the relation is quite loose and the resulted feature learner is often suboptimal in the sense of minimum Bayes error. In an independent work, Carneiro and Vasconcelos [4] proposed a joint feature extraction and selection algorithm by minimizing the Bayes error. For estimating the objective, the authors adopted a generative methodology by assuming Gaussian mixture models for the class-conditional distribution. A technical difficulty for this method is that the number of mixture components, which has dominant importance in their method, is very difficult to be determined in practice. In [32], Weston et al. proposed a learning algorithm that jointly learns feature selection and classification by minimizing a zero-norm regularized loss function to encourage featurewise sparseness of a classifier. To make the algorithm computation tractable, convex loss functions, e.g., the hinge loss function used in *support vector machines* (SVMs), have to be employed as the optimization objective. Although the Bayes error can be naturally reflected by the zero-one loss function, these convex surrogate loss functions offer poor approximation to the zero-one loss.

The recently proposed spectral framework [45] formulates feature selection as a graph-based optimization. Typically, a spectral feature learner is established by 1) conveying basic assumptions and heuristical intuitions into pairwise similarity of or constraints on the training instances; 2) constructing a affinity graph based on the defined similarity; and 3) building a learner by spectral analysis of the graph. Although the spectral graph theory [8] is usually employed to provide justifications for spectral selectors, it provides no guidelines on how to construct the graph, which is of central importance to the success of such graph-based learning algorithms since the performance is extremely sensitive to both graph structures and edge weight settings. As a consequence, one has to resort to heuristics to establish graph for spectral learning. In a recent work [38], we show that spectral feature learning algorithms could be derived from theoretic frameworks such as RoC and DoC. The resulted algorithms encode the RoC or DoC objectives into a well-defined graph and learns features by spectrally embedding the derived graph. The theoretic analysis and empirical results in [38] validate the advantage of our efforts in establishing feature learning algorithm top-down from a theoretic optimal perspective.

2.3 Feature Weighting and Relief

Feature weighting [33], [43], [5] is a greedy search strategy for feature selection. By assigning to each feature a real valued weight to indicate its quality, it simplify the combinatoric search problem to a continuous optimization task.

Among the existing feature weighting methods, the Relief algorithm [20], [26] is considered one of the most successful techniques due to its effectiveness and simplicity [25]. It is also frequently employed to enhance the performances of the lazy learning algorithms [33]. Recently, Gilad-Bachrach et al. [14] established a new variation of Relief based on the concept of margin. This idea was further exploited by Sun [28] to provide a new interpretation of Relief as a max-margin convex optimization problem. This perspective not only simplifies the computation of Relief significantly but also explains some property of Relief. However, the margin is a heuristical concept defined based on simple intuitions, thus, the secret behind the success of Relief is still unclear. In this paper, we show that DoC offers a more insightful interpretation to Relief.

According to Sun [28], Relief is equivalent to a convex optimization problem

$$\begin{aligned}&\max \sum_{n=1}^{N} \mathbf{w}^T\mathbf{m}_n\\&s.t.: ||\mathbf{w}||_2^2 = 1, \mathbf{w} \geq \mathbf{0},\end{aligned} \tag{3}$$

where $\mathbf{w} = (w_1, w_2, \ldots, w_D)^T$ is a weighting parameter vector, $\mathbf{m}_n = |\mathbf{x}_n - M(\mathbf{x}_n)| - |\mathbf{x}_n - H(\mathbf{x}_n)|$ is called the margin vector for the sample $\mathbf{x}_n$, $H(\mathbf{x}_n)$ and $M(\mathbf{x}_n)$ denote the nearest-hit (the nearest neighbor from the same class) and nearest-miss (the nearest neighbor from different classes) of $\mathbf{x}_n$, respectively. By using the Lagrangian relaxation and deriving the Karush-Kuhn-Tucker optimality, a simple close-form solution to (3) can be derived, i.e.,

$$\mathbf{w} = (\overline{\mathbf{m}})^+/||(\overline{\mathbf{m}})^+||_2, \tag{4}$$

where $\overline{\mathbf{m}} = \frac{1}{N}\sum_{n=1}^{n} \mathbf{m}_n$ is the averaged margin vector, $(\cdot)^+$ is an elementwise positive part operator, i.e., $(\mathbf{a})^+ = \max(\mathbf{a}, \mathbf{0})$.

3 A Discriminative Framework for Feature Selection

3.1 Discriminative Optimal Feature Selection

Feature selection plays a crucial role in machine learning. Yet, as oppose to learning algorithms, most of the feature selection criteria are based on heuristical intuitions or domain knowledge. A formal framework for feature selection in the literature is the representative optimal criteria framework proposed by Koller and Sahami [21], which placed a sound theoretic foundation for the feature selection task. Here, we propose another theoretic optimal framework as the discriminative counterpart for RoC. Unlike RoC, this criterion considers classification directly and naturally reflects the Bayes error rate in the reduced-dimensional space.

We view feature selection as a preprocessing procedure that is independent of any classifier [15], [21]. In other words, we focus on a Filtering setting. A discriminative approach would select a subset of M features such that the data are *essentially* most discriminant in terms of the selected features. For this purpose, assume δ is a decision rule (i.e., classifier) that maps each input sample $\mathbf{x}$ directly onto a class label y, the optimal feature selector should minimize the generalization error of an *ideal* classifier δ^*, which is, the Bayes error rate

$$\begin{aligned}&\min_{\tau} \inf_{\delta} E_{\mathbf{x}}[err(\delta|\tau(\mathbf{x}))]\\ &s.t.: \|\tau\|_0 = M,\end{aligned} \tag{5}$$

where $E_{\mathbf{x}}\{err(\delta|\tau(\mathbf{x}))\} = E_{\mathbf{x}}[1 - \max_c p(c|\tau(\mathbf{x}))]$ is the generalization error of a decision rule δ in the reduced-dimensional space, $c \in \{0, 1, \ldots, C-1\}$ is a class label, and $\inf$ denotes the infimum of a set. We call this optimal criterion as *discriminative optimal criterion* to highlight its goal to maximize the essential discriminative ability of the selected features.

The RoC criterion has been proved powerful for its ability in keeping maximum amount of information for modeling the posterior distribution $p(y|\tau(\mathbf{x}))$, which is of course useful for various domains [3], [21]. Besides, it is also closely related with the Bayes error for classification [17]. However, in the context of supervised classification, preserving the representative information to a very fine resolution might be wasteful of the limited training examples because our ultimate goal is merely to classify the data very coarsely into several classes. Moreover, for many applications where $\mathbf{x}$ have very high dimensionality, modeling the posterior probability with limited samples could be highly illposed. In contrast, there are plenty of compelling reasons for using discriminative criteria to assess the quality of features in classification scenarios [3]. One of such justifications is addressed concisely by Vapnik [30] as we quote: "one should solve the problem (classification) directly and never solve a more general problem (modeling $p(y|\mathbf{x})$) as an intermediate step."

3.2 Nonparametric Bayes Error Minimization

Both RoC and DoC are theoretically optimal but computationally intractable. In particular, we cannot compute the exact Bayes error in practice because neither the ideal decision rule nor the precise posterior distribution is known a priori. Therefore, approximation procedures for estimating the Bayes error is necessary. In general, there are two distinct approaches for estimating the Bayes error objective of DoC, which, interestingly, are reflected precisely into the taxonomy of feature selection algorithms.

- **Wrapper methods** optimizes the generalization error of a specific form decision rule $\delta \in \mathcal{H}$ through a nested/joint optimization

$$\begin{aligned}\tau &= \arg\min_{\tau} \min_{\delta\in\mathcal{H}} \{E_{\mathbf{x}}[err(\delta|\tau(\mathbf{x}))]\}\\ &= \arg\min_{\delta\in\mathcal{H},\tau} \{E_{\mathbf{x}}[err(\delta|\tau(\mathbf{x}))]\},\end{aligned} \tag{6}$$

  where $\mathcal{H}$ denotes a hypothesis space. Clearly, this is equivalent to estimate the Bayes risk by $\inf_{\delta\in\mathcal{H}}\{E_{\mathbf{x}}[err(\delta|\tau(\mathbf{x}))]\}$, which is then used as a measure to assess the usefulness of a feature subset so as to search for the optimal subset. However, because of the nontrivial coupling between τ and δ in the optimization, it is highly likely that the selected features are not optimal for classifiers other than the specified form (i.e., for decision rules not in the hypothesis space $\mathcal{H}$).
- **Filter methods**[1]estimates the Bayes risk without using a specific form of classifier, for instance, the approximation of the Bayes risk can be obtained by estimating the probability distribution involved in (5) [13], [14], or by a compact bound (upper or lower) of the Bayes risk [27], [35]. Here as an example, assume we have some kind of prior knowledge about the data and accordingly design a generative model. In particular, suppose we consider a binary-class Naïve Bayes generative model: $p(\mathbf{x}|y) = \prod_{d=1}^{D} p(x^{(d)}|y)$, where $y \in \{1, 0\}$, $p(x^{(d)}|y = c) = \mathcal{N}(x^{(d)}|\mu_{d,c}, \sigma_{d,c})$, the DoC framework naturally boils down to the following criterion for feature ranking:

$$\left(x_n^{(d)} - \mu_{d,1-y}\right)^2/\sigma_{d,1-y}^2 - \left(x_n^{(d)} - \mu_{d,y}\right)^2/\sigma_{d,y}^2.$$

  If we further assume identical variance $\sigma_{d,1}^2 = \sigma_{d,0}^2$ for the generative model, the criterion above turns out to be the feature ranking criterion well known as *Fisher Score* [18], [32]

$$FS_d = |\mu_{d,1} - \mu_{d,0}| \Big/ \sqrt{\sigma_{d,1}^2 + \sigma_{d,0}^2}.$$

[1] Note that filter methods could also be derived from the RoC framework, e.g., MI-based feature ranking.

In order to obtain generic approaches that are not confined to any learning algorithms, in this paper, we limit ourselves to the Filter setting for feature selection. Moreover, to establish feature learning algorithms that are robust to the underlying data distribution, we will focus on nonparametric techniques for estimating Bayes error, leading to a framework we referred to as nonparametric Bayes error minimization.

Given a set of training data $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$, the expectation over $\mathbf{x}$ can be approximated by the empirical average. To estimate $p(c|\tau(\mathbf{x}_n))$, there are two types of methods, namely, the parametric and nonparametric estimation methods. In this paper, we use nonparametric estimators to estimate $p(c|\tau(\mathbf{x}_n))$. Since we have

$$\begin{aligned} &E[1 - p(y|\tau(\mathbf{x}))] \\ &\quad = 1 - \frac{1}{2}E[p(y|\tau(\mathbf{x})) + 1 - p(c \neq y|\tau(\mathbf{x}))], \end{aligned}$$

we can minimize the Bayes error equivalently by optimizing the following objective:

$$\begin{aligned} J(\tau) &= E[p(y|\tau(\mathbf{x})) - \sum_{c \neq y} p(c|\tau(\mathbf{x}))] \\ &\propto \int_{\mathbf{x}} p(\mathbf{x})(p(y)p(\tau(\mathbf{x})|y) - \sum_{c \neq y} p(c)p(\tau(\mathbf{x})|c))d\mathbf{x} \\ &\approx \frac{1}{N}\sum_{n=1}^{N}\left(\hat{p}(y_n)\hat{p}(\tau(\mathbf{x}_n)|y_n) - \sum_{c \neq y_n}\hat{p}(c)\hat{p}(\tau(\mathbf{x}_n)|c)\right). \end{aligned} \tag{7}$$

While most classical parametric densities are unimodal (have a single peaks), nonparametric density models can effectively capture the multimodality properties that are frequently observed in many practical problems. Indeed, one of the advantages of using nonparametric estimators is its assumption-free property, i.e., the results will be robust to any probability distribution since the estimators do not rely on specific distribution assumptions. However, directly estimating $p(c|\tau(\mathbf{x}_n))$ based on the limited training data might incur singularity problems, because in most tasks, $\mathbf{x}_n$ is continuous and usually in a very high-dimensional space. Fortunately, because we always have $p(c|\mathbf{x}_n) \propto p(c)p(\mathbf{x}_n|c)$, we could adopt another much more efficient scheme. In particular, we can estimate $p(c)$ by the class ratio of the training data: $\sum_{n=1}^N I(y_n = c)/N$; thereafter, we can estimate $p(\mathbf{x}_n|c)$ efficiently using nonparametric estimators.

There are two commonly used nonparametric estimators, namely the k-nearest-neighbor method and the Parzen window method (or kernel density estimator). If a kNN estimator is employed, we have

$$\hat{p}(\mathbf{x}|c) = \frac{k/N_c}{V_c^{(k)}(x)}, \tag{8}$$

where N_c is the number of examples in the class with label c, and $V_c^{(k)}(\mathbf{x})$ denotes the volume of the hypersphere from $\mathbf{x}$ to its kth nearest neighbor in class c. Basically, this method employs a Monte Carlo procedure to estimate the probabilistic density by counting the proportion of the examples falling into the k-nearest-neighbor sphere.

In contrast, the Parzen-window estimator [3], [13] estimates the density function of a probability by averaging over all the examples with a kernel (or window) function. For example, to estimate $p(\mathbf{x}|c)$, we have

$$\hat{p}(\mathbf{x}|c) = \frac{1}{Z}\sum_{n:y_n=c} g\left(\frac{\mathbf{x} - \mathbf{x}_n}{\varsigma}\right), \tag{9}$$

where Z is a factor for normalization, $g(\cdot)$ is a kernel function with bandwidth parameter ς, for example, the popular Gaussian RBF kernel $\exp(-\frac{||\mathbf{x}-\mathbf{x}_n||^2}{2\varsigma^2})$.

So far, we have successfully established an algorithmic framework for feature selection based on the notion of nonparametric Bayes error minimization. Integrated with a chosen search strategy, a variety of feature selection algorithms can be derived from this framework. As a case study, we exemplify the framework in Section 4 by employ feature weighting as an example search strategy. We show that several existing feature weighting algorithms as well as new ones can be naturally derived from this framework.

4 Relief as Nonparametric Bayes Error Minimization

As we have mentioned, a feature selection algorithm consists of two necessary components, evaluation criterion and search strategy. We have successfully established a generic criterion for feature selection. To acquire a practical feature selection algorithm, we need to choose a specific search strategy. As a case study, in this section, we use feature weighting as search strategy for its simplicity and efficiency. However, extensions to other search strategies are also possible. For example, in [38], we established DoC-based algorithms for learning feature transformations, and in [39], we devise an algorithm that employ DoC for multiple kernel learning.

We first show that the famous Relief algorithm can be naturally derived from the proposed framework. Since Relief is originally established for binary classification tasks, we first consider binary labels, i.e., $y_n \in \{0, 1\}$. We will extend the results to multiclass scenarios later. Let us first assume a simplest case of balanced class distribution $p(c = 1) = p(c = 0) = 0.5$, and estimate the class-conditional probability by a simple 1-Nearest Neighbor (1NN) estimator, we have

$$\begin{aligned} \hat{J}_n &= \hat{p}(c = y_n|\tau(\mathbf{x}_n)) - \hat{p}(c \neq y_n|\tau(\mathbf{x}_n)) \\ &\propto \hat{p}(\tau(\mathbf{x}_n)|c = y_n) - \hat{p}(\tau(\mathbf{x}_n)|c \neq y_n) \\ &= \frac{1/N}{V^{(1)}} - \frac{1/N}{V^{(2)}} \\ &\propto \frac{1}{||\tau(\mathbf{x}_n) - H(\tau(\mathbf{x}_n))||_2^M} - \frac{1}{||\tau(\mathbf{x}_n) - M(\tau(\mathbf{x}_n))||_2^M}, \end{aligned} \tag{10}$$

where J_n denotes the objective w.r.t. the sample $\mathbf{x}_n$, $H(\mathbf{x}_n)$ and $M(\mathbf{x}_n)$ are the nearest-hit and nearest-miss of $\mathbf{x}_n$, respectively, $V^{(1)}$ and $V^{(2)}$ denote the volumes of the hyperspheres from $\mathbf{x}_n$ to $H(\mathbf{x}_n)$ and to $M(\mathbf{x}_n)$, respectively. We then plug this evaluation criterion (10) into the feature weighting optimization scheme [28]. Basically, feature weighting assigns a real-valued weight for each feature and approximates the goodness of a feature set by a linear combination of the goodness of each feature $J(\tau) = \sum_{\tau_d=1} w_d J^{(d)}$, where $J^{(d)}$

denotes the quality measure of the dth feature, w_d is the weight for the dth feature $x^{(d)}$. We have

$$\begin{aligned} &\max \sum_{d=1}^{D} w_d \sum_{n=1}^{N} \hat{J}_n^{(d)} \\ &s.t. : ||\mathbf{w}||_2^2 = 1, \mathbf{w} \geq \mathbf{0}, \end{aligned} \tag{11}$$

where

$$\begin{aligned} \hat{J}_n^{(d)} &= \hat{p}\big(c = y_n | x_n^{(d)}\big) - \hat{p}\big(c \neq y_n | x_n^{(d)}\big) \\ &= \frac{1}{\big|x_n^{(d)} - H(x_n^{(d)})\big|} - \frac{1}{\big|x_n^{(d)} - M(x_n^{(d)})\big|}, \end{aligned}$$

$M_n^{(d)}$ and $H_n^{(d)}$ denote the nearest-miss and nearest-hit of $\mathbf{x}_n$ in the dth dimensional subspace.

Clearly, (11) will be identical to the Relief optimization of (3) if we further take two approximation precedures: 1) use $J_n^{(d)} = |x_n^{(d)} - M_n^{(d)}| - |x_n^{(d)} - H_n^{(d)}|$, a loosely equivalent term to optimize [18], and 2) approximate $M_n^{(d)}$ and $H_n^{(d)}$ with $M^{(d)}(\mathbf{x}_n)$ and $H^{(d)}(\mathbf{x}_n)$, which denote the dth elements of $M(\mathbf{x}_n)$ and $H(\mathbf{x}_n)$, respectively. The second approximation means, instead of using D 1D nearest-miss's $M_n^{(d)}$, we use a single D-dimensional nearest-miss $M(\mathbf{x}_n)$ and approximate the dth 1D nearest-miss $M_n^{(d)}$ with the dth element of $M(\mathbf{x}_n)$. Note that the second approximation could be very poor if the data contains a large number of redundant or noisy features. For a discussion on this issue, the readers may refer to [37].

We summarize the findings into a new interpretation to the Relief Algorithm.

Proposition 4.1. *The Relief algorithm approximates a feature selection method based on nonparametric Bayes error minimization with assumptions:*

1. *Binary classification task.*
2. *Balanced label distribution.*
3. *1NN estimator.*
4. *Greedy search based on feature weighting.*

RELIEF was originally proposed as an online feature selection algorithm based on some heuristical intuitions [20]. It is considered one of the most successful algorithms for assessing the quality of features. As the first effort toward the better understanding of this algorithm, Robnik-Sikonja and Kononenko [25] interpreted Relief's evaluation criterion as the ability to explain particular concept, i.e., "the ratio between the number of the explained changes in the concept and the number of the examined instances." Recently, Sun [28] proved that RELIEF is an online solution to a convex optimization problem, where the evaluation criterion is a margin-based loss function. While this interpretation is able to explain some of the properties of Relief, the success of these max-margin feature selectors is rarely investigated. Particularly since the margin is a heuristical concept, behind which the secret is still at large. Here, by deriving from a theoretically optimal framework, we offer a new interpretation to Relief as well as to other margin-based feature weighting algorithms. This new perspective raises opportunities to establish new max-margin feature weighting alternatives and also to identify weakness of existing feature weighting methods so as to improve them.

4.1 Parzen-Relief

By exploiting the new interpretation of Relief, we establish an alternative feature weighting algorithm which resembles the standard Relief algorithm in all aspects except that the Parzen widow instead of kNN estimator is used to estimate the conditional distribution.

We use the well-known isotropic Gaussian RBF kernel

$$k(\mathbf{x}, \mathbf{x}') = \frac{1}{\sqrt{2\pi}\varsigma} \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||}{2\varsigma^2}\right). \tag{12}$$

Given a set of training data $\{\mathbf{x}_n\}_{n=1}^{N}, p(\mathbf{x})$ can be estimated by

$$\hat{p}(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n).$$

Plugging the above into the objective (7), we have

$$\max \sum_{n=1}^{N} \left(\frac{1}{|\Omega_n^{(o)}|} \sum_{i \in \Omega_n^{(o)}} k(\mathbf{x}_n, \mathbf{x}_i) - \frac{1}{|\Omega_n^{(e)}|} \sum_{j \in \Omega_n^{(e)}} k(\mathbf{x}_n, \mathbf{x}_j) \right), \tag{13}$$

where $\Omega_n^{(o)} = \{i : y_i = y_n\}$ and $\Omega_n^{(e)} = \{j : y_j \neq y_n\}$ denote the Homogenous and Heterogeneous Index Set of $\mathbf{x}_n$.

Using feature weighting as search strategy, which optimizes the quality of the feature set by greedily optimizing the weighted linear combination of the goodness of each feature, we have

$$\mathbf{w} = \mathop{\arg\max}_{\mathbf{w} \geq \mathbf{0}, ||\mathbf{w}||_2^2 = 1} \sum_{d=1}^{D} w_d \sum_{n=1}^{N} \hat{p}\big(c = y_n | x_n^{(d)}\big) - \hat{p}\big(c \neq y_n | x_n^{(d)}\big).$$

Following a series of similar simplification, we get a new feature weighting method, which we called *"Parzen-Relief"*

$$\begin{aligned} &\max \sum_{n=1}^{N} \mathbf{w}^T \mathbf{m}_n^p \\ &s.t. : ||\mathbf{w}||_2^2 = 1, \mathbf{w} \geq \mathbf{0}, \end{aligned} \tag{14}$$

where the margin $\mathbf{m}_n^p = [m_n^{p(1)}, \ldots, m_n^{p(D)}]^T$, the dth element is defined as $m_n^{p(d)}$=:

$$\frac{1}{|\Omega_n^{(o)}|} \sum_{i \in \Omega_n^{(o)}} k_d\big(x_i^{(d)}, x_n^{(d)}\big) - \frac{1}{|\Omega_n^{(e)}|} \sum_{j \in \Omega_n^{(e)}} k_d\big(x_j^{(d)}, x_n^{(d)}\big),$$

and $k_d(x_1^{(d)}, x_2^{(d)}) = \exp(-\frac{|x_1^{(d)} - x_2^{(d)}|^2}{2\varsigma_d^2})$.

Note that the P-Relief algorithm has the same computational complexity as the standard Relief algorithm. The only difference between Relief and P-relief is the step computing the margin, in which Relief looks through the homogenous and heterogenous index set to find the nearest-hit and nearest-miss for each $\mathbf{X}_n$, while P-Relief averaging the kernel terms $k(\mathbf{x}_i, \mathbf{x}_n)$ for $\mathbf{x}_i$ in the homogenous set and $k(\mathbf{x}_j, \mathbf{x}_n)$ for $\mathbf{x}_j$ in heterogenous set. Clearly, these two steps are of the same complexity, $O(|\Omega_n^{(o)}| + |\Omega_n^{(e)}|)$.

4.2 MAP-Relief

An undesirable assumption made by the Relief algorithm, as suggested by our derivation of Proposition 4.1, is that the label distribution is balanced among "one-versus-rest" split of different classes such that to maximize a posterior (*MAP*) probability is equivalent to maximize the class conditional probability (likelihood). In multiclass classification scenarios, this implies $P(c) = 1/2$ for all the classes $c = 0, 1, \ldots, C$, which is clearly impossible in practice. To address this problem, we derive from our proposed algorithmic framework to get a refined definition of the margin which turns out to integrate the label distribution into the max-margin objective. In particular, we have

$$\begin{aligned}&\hat{p}(c = y_n|\tau(\mathbf{x}_n)) - \hat{p}(c \neq y_n|\tau(\mathbf{x}_n))\\ &\quad\propto \hat{p}(c = y_n)\hat{p}(\tau(\mathbf{x}_n)|c = y_n) - \hat{p}(c \neq y_n)\hat{p}(\tau(\mathbf{x}_n)|c \neq y_n)\\ &\quad\propto \frac{\hat{p}(y_n)}{||\tau(\mathbf{x}_n) - H(\tau(\mathbf{x}_n))||_2^M} - \frac{(1 - \hat{p}(y_n))}{||\tau(\mathbf{x}_n) - M(\tau(\mathbf{x}_n))||_2^M},\end{aligned} \tag{15}$$

Incorporating this objective into the feature weighting optimization formulation, we obtain the following convex optimization problem:

$$\begin{aligned}&\max \sum_{n=1}^{N} \mathbf{w}^T \mathbf{m}_n^{\pi}\\ &s.t.: ||\mathbf{w}||_2^2 = 1, \mathbf{w} \geq \mathbf{0},\end{aligned} \tag{16}$$

where the refined margin is given by

$$\mathbf{m}_n^{\pi} = \hat{p}(y_n)|\mathbf{x}_n - M(\mathbf{x}_n)| - (1 - \hat{p}(y_n))|\mathbf{x}_n - H(\mathbf{x}_n)|,$$

$\pi = [P(0), \ldots, P(C-1)]$ is the label distribution. We term this algorithm as *MAP-Relief*, where MAP is an abbreviation of *max a posterior*. In Section 5, we will show that while the performance of other algorithms in the Relief family degrades significantly when the data set is strongly imbalanced, M-Relief is much more robust.

Besides its robustness in handling imbalanced data, another advantage of M-Relief algorithm is that it can naturally deal with multi class tasks. The original Relief algorithm only works for binary classification problems [20]. *ReliefF* [22] extends it to multiclass scenarios by a heuristic updating rule, which is equivalent to solve Relief with the margin vector

$$\mathbf{m}_n^F = \sum_{c \neq y_n} \hat{p}(c)|\mathbf{x}_n - M_c(\mathbf{x}_n)| - |\mathbf{x}_n - H(\mathbf{x}_n)|,$$

where $M_c(\mathbf{x}_n)$ is the nearest miss of $\mathbf{x}_n$ in class c, $c \in 0, \ldots, C-1$. Therefore, one needs to search for k-nearest-hit and $k(C-1)$-nearest-miss for each sample to solve ReliefF. However, from our new interpretation of Relief, this is clearly unnecessary, because in general the following relationship always holds:

$$\sum_{c \neq y_n} p(c)p(\mathbf{x}_n|c) = p(\mathbf{x}_n, c \neq y_n) = (1 - p(y_n))p(\mathbf{x}_n|c \neq y_n).$$

The Iterative-Relief (I-Relief, [28]) algorithm deals with multiclass data using a margin vector defined somewhat similar with our definition $\mathbf{m}_n^{\pi}$, but with an implicit assumption $P(y_n) = 0.5$. Obviously, this assumption is inappropriate for problems involving $(C > 3)$ categories. This could become more severe when C goes larger such that the "one-versus-rest" splits (i.e., $\{\mathbf{x}_i : y_i = c\}$ and $\{\mathbf{x}_j: y_j \neq c\}$) of the data set become more and more imbalanced.

TABLE 1
Characteristics of 12 UCI Data Sets

Data Set	Train Size	Test Size	#Feature	#Class
Breast	400	283	9	2
German	700	300	20	2
Ionosphere	235	116	34	2
Waveform	400	4600	21	2
Pima	400	368	8	2
Heart	170	100	13	2
Sonar	165	43	60	2
Splice	1000	2175	60	2
LRS	380	151	93	48
Glass	120	94	9	6
Ecoli	200	136	7	8
Segment	210	2100	18	7

It is interesting to see that M-Relief possesses advantages of both ReliefF and I-Relief, and at the same time mitigates their drawbacks: 1) Similar with ReliefF, M-Relief incorporates the class distribution to tackling imbalanceness; 2) Similar with I-Relief, M-Relief needs only one, instead of $k \times (C-1)$, nearest-miss for each pattern. Both advantages, i.e., computational efficiency and ability to handling imbalanced data, would become especially valuable when we are facing problems with very large number of classes.

5 Experiments

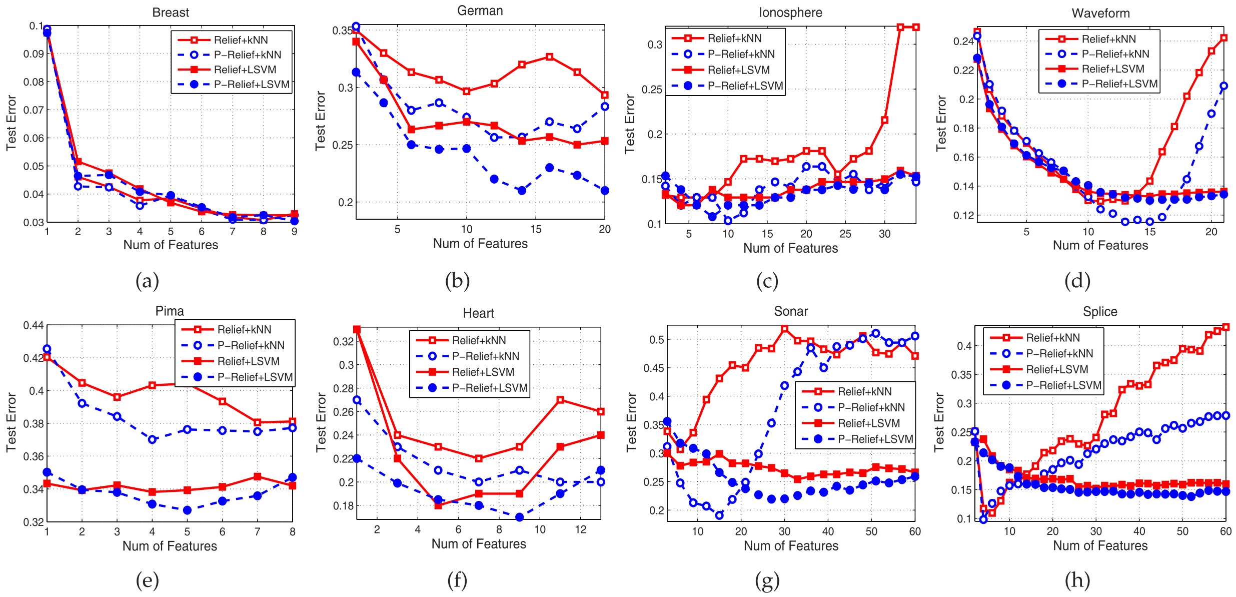
In this section, we empirically validate our claims by conducting extensive experiments to evaluate the effectiveness of the proposed methods. To achieve fair comparisons, we only consider algorithms with greedy search strategies such as feature weighting or feature ranking as baselines. We first test the proposed algorithms along with other competitors on UCI data sets in a controlled manner. We then extend our experiments with applications to two large-scale real-world tasks, i.e., text term selection and microarray gene selection.

5.1 Experiments on UCI Data Sets

To demonstrate the performance and empirical behavior of the proposed feature weighting algorithms, we first conducted experiments in a controlled manner. For this purpose, 12 benchmark machine learning data sets from the UCI machine learning repository[2] are selected because of their diversity in the numbers of features, instances, and classes. The statistical information of the data sets are summarized in Table 1. To establish a controllable experiment setting and also to facilitate the comparison, 50 irrelevant features (known as "probes") are added to each data example. Each of the probes is sampled independently from a spherical Gaussian distribution $\mathcal{N}(0, 20)$.[3] Two distinct metrics are used to evaluate the effectiveness of the feature selection algorithms. One is the classification accuracy of a kNN classifier (in some cases also by the Lagrangian Support Vector Machine (LSVM, [23]), an efficient implementation of SVMs). The other metric is the Receiver

[2] Available at: http://archive.ics.uci.edu/ml/.

[3] Experimental results with probes generated from other distributions (e.g., uniform, multinomial) are similar.



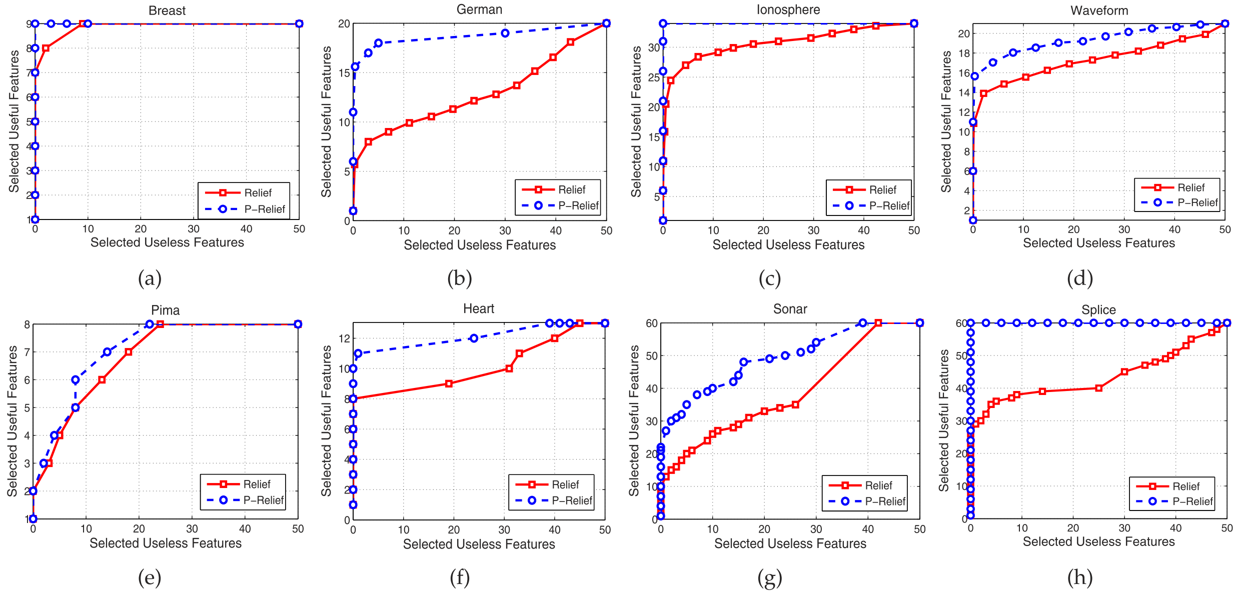
(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 1. Comparison of P-Relief and Relief on UCI data: testing accuracy versus the number of selected features.

Operating Characteristic (ROC) curve [28], which can indicate the effectiveness of different feature selection algorithms in identifying relevant features while ruling out useless ones. To eliminate statistical deviations, all the experiments are averaged over 20 random runs. The hyperparameters, for example, the number of nearest neighbors k in kNN and the regularization parameter C in LSVM are tuned by a fivefold cross validation procedure using the training data.

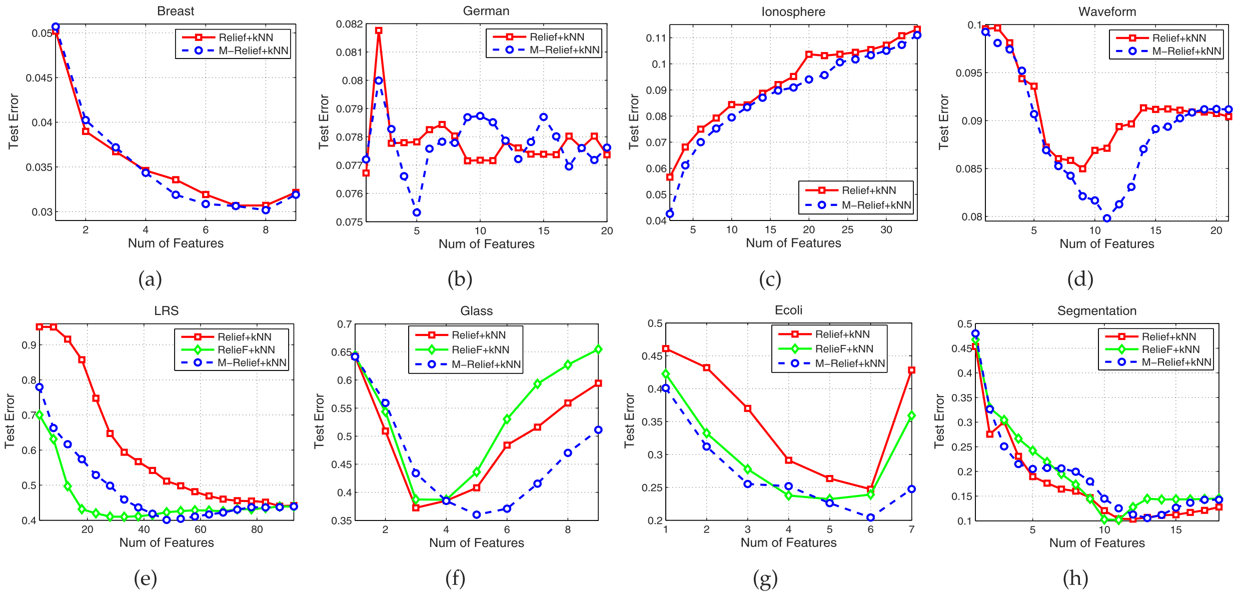
We first apply Relief and P-Relief to the eight binary classification data sets. The purpose of this experiment is to examine which nonparametric estimator (kNN or Parzen) works better for the nonparametric DoC framework. For this comparison, both kNN and LSVM are used as the classifier. The hyperparameter in P-Relief, i.e., the kernel (widow) width) is fixed at a default value: $\varsigma_d = 0.01$. Fig. 1 shows the average testing error of each selector-classifier combination, as a function of the number of top-ranked features. The ROC curves are plotted in the Fig. 2. As a reference, the best average classification error and standard deviation of each algorithm are also plotted as a bar chart in Fig. 5. From these results, we can see that, although P-Relief shares the same computational complexity as Relief, it usually achieves better performance than Relief. In particular, in terms of the testing set classification accuracy, P-Relief outperforms Relief in seven (out of eight) data sets and performs comparably on the other one. The improvements on the seven data sets are all significant, according to student t-test with confidence threshold 0.01. These observations suggest that, for estimating the DoC criterion for feature weighting, Parzen-window approach seems preferable over kNN estimator as it empirically leads to better



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 2. Comparison of P-Relief and Relief on UCI data: feature selection ROC.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 3. Comparison of Relief, ReliefF, and M-Relief on UCI data: testing accuracy versus feature set size.

classification accuracy. With respect to the ROC measure, we see that, for all the eight data sets, P-Relief achieves a much larger area under ROC curve than Relief does. For some data sets (e.g., Breast, Ionosphere, Splice), P-Relief even achieves an ideal ROC shape. These observations suggest that the P-Relief has a better ability of retaining useful features while eliminating useless ones, and further indicates that, the Parzen-window method can achieve better estimations of the DoC-based feature quality.

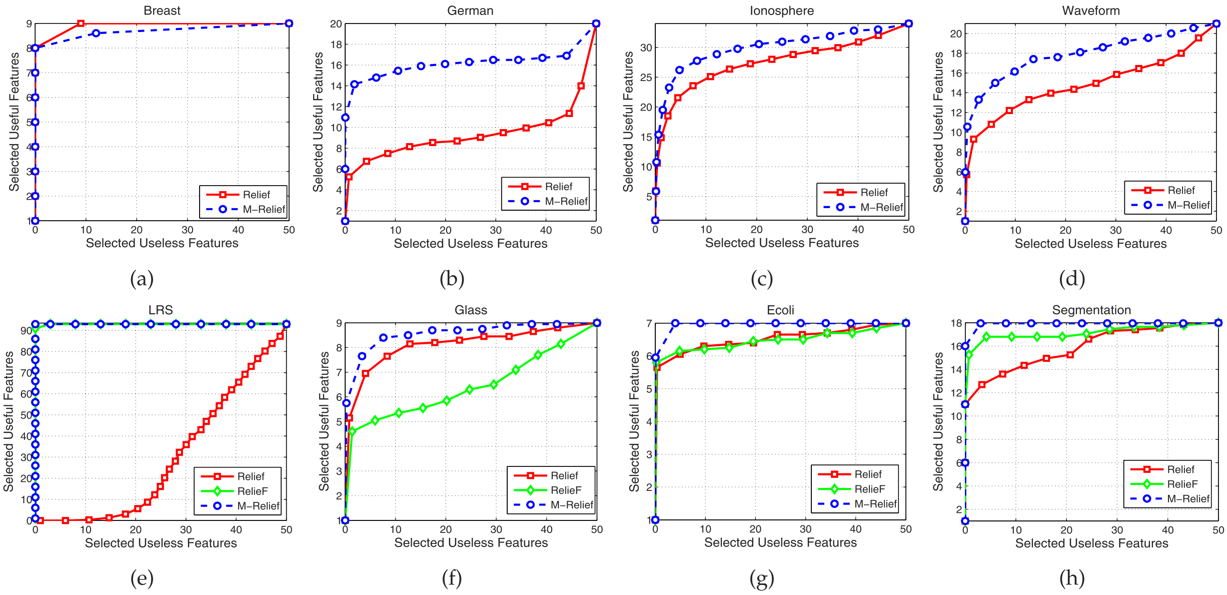
One of our contributions is that we identified the weakness of Relief in handling imbalanced data and accordingly proposed the MAP-Relief algorithm to address the issue. We now test this claim by comparing Relief, ReliefF and the proposed M-Relief on imbalanced and/or multiclass data. For this purpose, four binary data sets, which are relatively more imbalanced, and four multiclass data sets are used. To facilitate the comparison, a biased-sampling procedure is further applied to the four binary classification data sets to make them even more imbalanced. Particularly, we randomly sampled 80 percent of the sample data from the minority class and hid them in the experiment. Since Relief is originally established for binary classification, to enable it to apply to multiclass tasks, we use the margin definition used in I-Relief [28]. To achieve a fair comparison, we use one nearest hit and $C-1$ nearest misses (one for each class) in ReliefF. This configuration is to ensure that the performance differences are mainly caused by the strategies used to handling imbalance rather than other factors.

Among the three algorithms, ReliefF is relatively more time-consuming than the other two because it needs searching nearest miss for each class. However, the differences are not very significant because the number of classes are not very large. For convenience, only the kNN classifier is used to estimate the classification error. Fig. 3 shows the testing error of each approach, as a function of the number of top-ranked features. The ROC curves are plotted in Fig. 4. And the bar plot indicating the best average testing errors and standard deviations are also shown in Fig. 5. Note that for binary classification, ReliefF and Relief are identical to each other. From these results, we arrive at the following observations: 1 the performance of Relief is degraded significantly when the data is highly imbalanced; 2) By exploiting the proposed framework to integrate the label distribution with the margin definition, M-Relief improves the performance significantly while not introducing much extra computation. In particular, it performs the best in six (out of eight) data sets in terms of the classification error metric, and in seven data sets in terms of the ROC metric. According to a similar significance test, M-Relief significantly improves the classification accuracy on six out of the eight data sets.

5.2 Term Selection for Natural Language Text Classification

An important application of feature selection is to select the most informative terms (words or phrases) for natural language text classification. This task is important yet challenging as the original feature space (vocabulary) usually consists of hundreds of thousands of terms. The extremely high dimensionality of natural language texts is a core challenge for text classification, and term selection has been shown to be the most effective way to improve both classification performance and computational efficiency [41]. In this section, we apply our algorithms to this task. Six benchmark text data sets from Trec (the Text REtrieval Contest, http://trec.nist.gov) collection that are frequently used in natural language analysis are selected. The information of each data set is summarized in Table 2. Note that, to illustrate the degree of imbalance of each data set, the $\max_c P(c)$ is also given in the last column of Table 2.

We compare ReliefF and P-Relief on these data sets with no probe added. For evaluation, k-NN classifiers are applied to the reduced-dimensional data, and the Macro-average F_1 ($Macro_{ave}\mathrm{F}_1$) and Micro-average F_1 ($Micro_{ave}\mathrm{F}_1$) [40] are used to assess the classification results. To obtain statistically reliable results, tenfold cross validation (i.e., 90 percent as



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 4. Comparison of Relief, ReliefF, and M-Relief on UCI data: feature selection ROC.

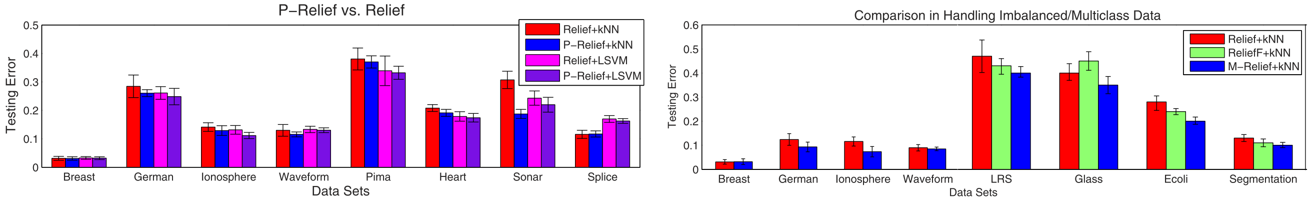


Fig. 5. Average testing errors and standard deviations, at optimal feature size.

training data and 10 percent as testing data for 10 repetitions) is used to estimate these evaluation metrics.

Fig. 6 plots the performance curves (i.e., accuracies versus number of selected terms) of ReliefF and P-Relief on the six text data sets. From the results, we can see that P-Relief outperforms ReliefF drastically in almost all the data sets. In addition, our algorithm is able to eliminate up to 98 percent of terms while improving the classification performance by up to 0.38 (relative: 85 percent), compared to the results of ReliefF or those without term selection. As a reference, we also show in Fig. 7 the amounts of performance improvements of P-Relief compared to ReliefF and all-features (i.e., results without feature selection).

Another interesting observation is that the scores learned by our algorithm are intuitively more reasonable than those learned by ReliefF. As shown in Fig. 8, P-Relief scores naturally in favor of common terms over rare terms, which is consistent with a well-known heuristic in natural language analysis that common terms (i.e., terms with higher DF or TF) are usually more informative than rare ones for text classification [41]. In contrast, ReliefF occasionally assigns high credits for terms that appear only once in the whole corpus.

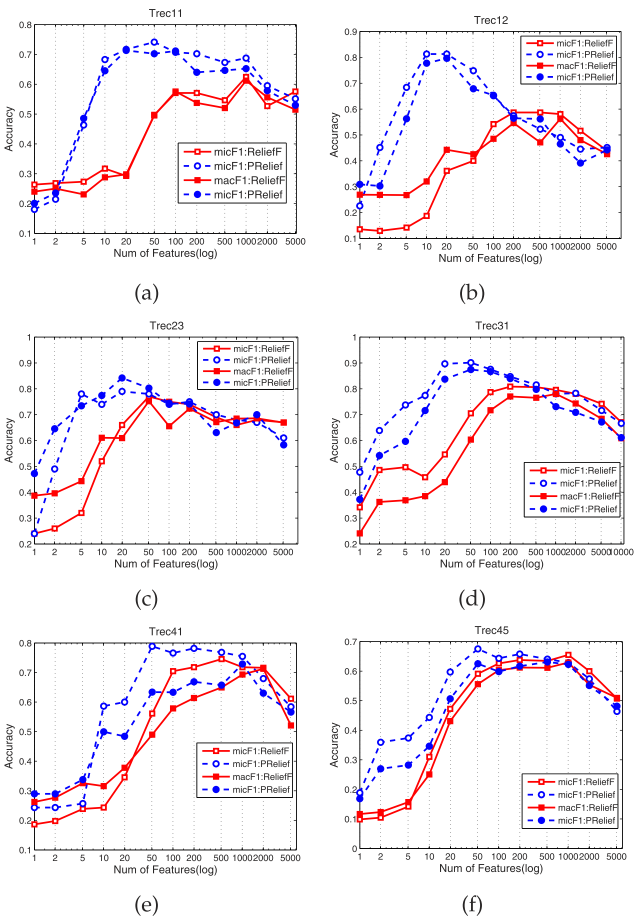
TABLE 2
Statistics of Six Natural Language Text Data Sets and Six DNA Microarray Gene Expression Data Sets

Data Set	#Sample	#Feature	#F.per S.	#Class	Max P_c
trec11	414	6424	16	9	31.8%
trec12	313	5799	19	8	29.7%
trec23	204	5831	29	6	44.6%
trec31	927	10127	11	7	37.9%
trec41	878	7453	8	10	27.7%
trec45	690	8261	12	10	23.2%
9tumors	60	5726	95	9	15.0%
leukemia1	72	5327	74	3	52.8%
leukemia2	72	11225	156	3	38.9%
lungCancer	203	12600	62	5	68.5%
SRBCT	83	2308	28	4	34.9%
DLBCL	77	5469	71	2	75.3%

5.3 Gene Selection for Microarray Genomic Cancer Diagnosis

We finally apply our algorithms to the task of gene selection for cancer diagnosis from DNA microarray gene expression data. For this purpose, we select six well-known microarray data sets, `9 Tumors`, `Leukemia1`, `Leukemia2`, `Lung Cancer`, `SRBCT`, and `DLBCLA`, which are all publicly available at http://www.gems-system.org. The information of each data set is also given in Table 2.

A practical challenge in gene selection is that the microarray data sets usually consist of a very limited number of samples (usually less than 100) with extremely high dimensionality (usually greater than 10,000). As shown in Table 2, the feature-size to sample-size ratio is as high as 156 (i.e., the number of features are 156 times as many as the number of examples). This severe curse of dimensionality is a key challenge in genomic microarray data analysis, e.g., a simplest linear learner could be highly underdetermined on such data sets. Therefore, for this task, both classification accuracy and the size of selected features are crucially important, that is, we want to achieve as high as possible accuracy with as few as possible features.



(a) (b) (c) (d) (e) (f)

Fig. 6. Average F_1-measures of ReliefF and P-Relief on six text classification data sets.

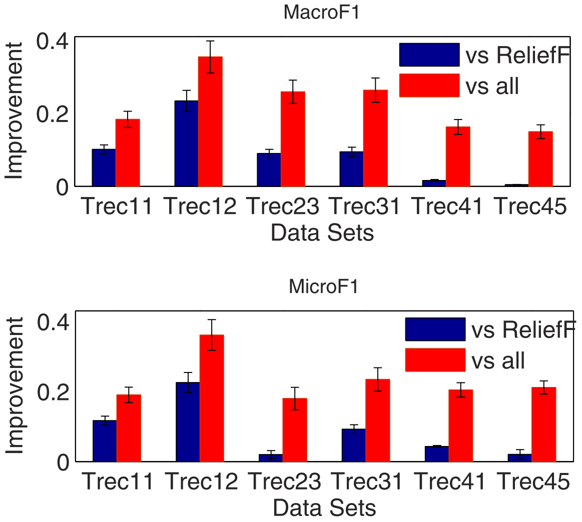


Fig. 7. Average improvements and standard deviations of P-Relief over ReliefF and all-features on text data sets.

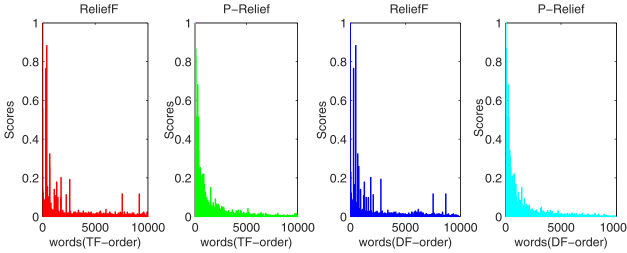


Fig. 8. Learned feature (word) weights of ReliefF and P-Relief ordered by term frequency (TF) or document frequency (DF).

We compare the two DoC-based algorithms, ReliefF and P-Relief, and two famous RoC-based feature ranking algorithm, Information Gain (IG) and Chi-Square (χ^2) [41], [34], on these data sets. According to existing reports [34], [41], [16], these two RoC approaches are among the best performing feature rankers on microarray data. Similar to the term selection experiment, because the feature sizes of microarray are already very large, we do not add any probes. For testing, we apply k-NN classifiers, and $Macro_{ave}F_1$ and $Micro_{ave}F_1$ are used as evaluation metrics. Due to the sparsity of data, all the results are obtained by using leave-one-out cross validation. For the sake of computational feasibility, each algorithm is only evaluated at feature size: 1, 2, 5, 10, 20, 40, 60, 100, 200, 500, and ALL. As we already mentioned that both accuracy and effective feature size are important for this task, we report both the classification accuracy and the best feature size in Table 3. From Table 3, we can see that for almost all the data sets, the two DoC-based approaches, ReliefF and P-Relief, perform significantly better than the two RoC-based approaches (IG and χ^2). Both ReliefF and P-Relief substantially improve the performance of k-NN compared to using all genes, whereas IG and χ^2 perform slightly worse on some of the data sets (e.g., leukemia2). This observation empirically validates the effectiveness of our DoC framework in comparison with the RoC framework. In addition, we can see that, except for 9-Tumors, in which our proposed algorithm P-Relief performs comparably with ReliefF, for all the other five data sets, P-Relief consistently outperforms ReliefF. For example, for leukemia1, SRBCT, and DLBCT, our algorithm is able to achieve 100 percent accuracy with only tens of genes. Although ReliefF achieves similar performance on

TABLE 3
Comparison of ReliefF, P-Relief, and Information Gain on Microarray Gene Expression Data: Each Entry Represents the Average F_1 Measure at the Best Feature Size, the Number in () Represents the Corresponding Feature Size

Data Set	$\mathrm{Macro}_{avg}\mathrm{F1}$					$\mathrm{Micro}_{avg}\mathrm{F1}$				
	All	ReliefF	PRelief	IG	χ^2	All	ReliefF	PRelief	IG	χ^2
9tumors	.67(5726)	**.87**(200)	.85(60)	.74(500)	.84(500)	.50(5726)	**.86**(200)	.84(100)	.53(500)	.80(500)
leukemia1	.89(5327)	.99(40)	**1.00**(40)	.94(60)	.91(100)	.91(5327)	.97(40)	**1.00**(40)	.91(40)	.87(100)
leukemia2	.97(11225)	.96(500)	**.99**(500)	.97(500)	.97(500)	.94(11225)	.94(40)	**.97**(500)	.94(500)	.94(500)
lungCancer	.82(12600)	.91(500)	**.95**(200)	.87(60)	.91(500)	.90(12600)	.93(500)	**.96**(500)	.89(500)	.86(200)
SRBCT	.77(2308)	1.00(100)	**1.00(10)**	.89(40)	.92(100)	.72(2308)	1.00(100)	**1.00(10)**	.80(40)	.90(100)
DLBCL	.96(5469)	1.00(100)	**1.00(60)**	.92(500)	.91(200)	.90(5469)	1.00(100)	**1.00(60)**	.85(200)	.87(200)

The best results are highlighted in bold.

SRBCT and DLBCT, the number of genes selected by ReliefF are relatively larger than P-Relief. Indeed, the high accuracy and compact gene expression obtained by our algorithm would enable experts in bioinformatics look directly into these genes to infer the molecular mechanisms and underlying causes of these cancers.

6 Conclusion

A natural criterion for feature selection would be to minimize the Bayes error in the reduced-dimensional space, because the generalization error of any classifier is lower bounded by Bayes error and the Bayes error only depends on features rather than classifiers, hence, Bayes error serves as an ideal measure to assess the quality of feature subsets. Based on this notion, we have presented a discriminative optimal criterion for feature selection, which possesses several compelling merits compared with its representative counterpart.

Although theoretically optimal, the discriminative optimal criterion is computationally intractable as it involves probabilities that are not known a priori. To this end, we have presented an algorithmic framework for feature selection based on nonparametric Bayes error minimization. We show that the proposed framework offers sound interpretations to existing approaches and also provide principled building blocks for establishing new algorithms. For example, when feature weighting are used as the search strategy, this framework reveals that the Relief algorithm greedily attempts to minimize Bayes error estimated by kNN estimator. The new interpretation of Relief insightfully explains the secret behind the heuristical margin. As an alternative, a new algorithm named Parzen-Relief is proposed. Furthermore, the new interpretation enables us to identify the weaknesses of Relief so as to improve it. In particular, to enhance its ability in dealing with imbalanced and/or multiclass data, we have proposed a MAP-Relief algorithm, which exploits the proposed framework to take advantage of the label distributions, leading to an weighted max-margin optimization problem.

Being independent to any feature subset search strategy, the proposed algorithmic framework is generic enough for establishing various feature selection algorithms. In [38], we showed that DoC can be employed for learning affinity graphs or pairwise similarities and developed an algorithm for learning feature transformation. In [39], we further exploited the DoC framework and established efficient algorithms for multiple kernel learning as well as ranking aggregation. For more systematic study, we plan to investigate new DoC-based algorithm instances for solving real-world challenges. We would also like to examine the asymptotic properties [2] of DoC and establish theoretic guarantee for DoC-based algorithms.

Acknowledgments

This work is based on the preliminary results presented in the PAKDD awarded paper [36]. Part of the work was support by a grant from Natural Sciences Foundation of China (#60275025).

References

[1] M. Belkin and P. Niyogi, “Laplacian Eigenmaps for Dimensionality Reduction and Data Representation,” *Neural Computation,* vol. 15, no. 6, pp. 1373-1396, June 2003.

[2] P.J. Bickel, Y. Ritov, and A. Tsybakov, “Simultaneous Analysis of Lasso and Dantzig Selector,” *Annals of Statistics,* vol. 37, no. 4, pp. 1705-1732, 2009.

[3] C.M. Bishop, *Pattern Recognition and Machine Learning,* Information Science and Statistics, 1 ed. Springer, 2007.

[4] G. Carneiro and N. Vasconcelos, “Minimum Bayes Error Features for Visual Recognition by Sequential Feature Selection and Extraction,” *Proc. Computer and Robot Vision Conf. (CRV ’05),* pp. 253-260, 2005.

[5] B. Chen, H. Liu, J. Chai, and Z. Bao, “Large Margin Feature Weighting Method via Linear Programming,” *IEEE Trans. Knowledge Data Eng.,* vol. 21, no. 10, pp. 1475-1488, Oct. 2009.

[6] Q. Chen and Y.-P.P. Chen, “Discovery of Structural and Functional Features in RNA Pseudoknots,” *IEEE Trans. Knowledge Data Eng.,* vol. 21, no. 7, pp. 974-984, July 2009.

[7] E. Choi and C. Lee, “Feature Extraction Based on the Bhattacharyya Distance,” *Pattern Recognition,* vol. 36, no. 8, pp. 1703-1709, 2003.

[8] F.R.K. Chung, “Spectral Graph Theory,” *Proc. Regional Conf. in Math. (AMS 1992),* vol. 92, 1997.

[9] E.F. Combarro, E. Montañés, I. Díaz, J. Ranilla, and R. Mones, “Introducing a Family of Linear Measures for Feature Selection in Text Categorization,” *IEEE Trans. Knowledge Data Eng.,* vol. 17, no. 9, pp. 1223-1232, Sept. 2005.

[10] M. Dash and H. Liu, “Feature Selection for Classification,” *Intelligent Data Analysis,* vol. 1, nos. 1-4, pp. 131-156, 1997.

[11] R. Duda, P.E. Hart, and D.G. Stock, *Pattern Classification,* ch. 4, second, ed., pp. 161-214, John-Wiley & Sons, 2001.

[12] J. Fan and R. Li, “Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery,” *Proc. Madrid Int’l Congress of Mathematicians (ICM ’06),* pp. 595-622, 2006.

[13] K. Fukunaga and D.M. Hummels, “Bayes Error Estimation Using Parzen and k-NN Procedures,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. PAMI-9, no. 5, pp. 634-643, Sept. 1987.

[14] R. Gilad-Bachrach, A. Navot, and N. Tishby, “Margin Based Feature Selection—Theory and Algorithms,” *Proc. 21st Int’l Conf. Machine Learning (ICML ’04),* 2004.

[15] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” *J. Machine Learning Research,* vol. 3, pp. 1157-1182, 2003.

[16] M.A. Hall and G. Holmes, “Benchmarking Attribute Selection Techniques for Discrete Class Data Mining,” *IEEE Trans. Knowledge and Data Eng.,* vol. 15, no. 6, pp. 1437-1447, Nov./Dec. 2003.

[17] K.E. Hild, D. Erdogmus, K. Torkkola, and J.C. Principe, “Feature Extraction Using Information-Theoretic Learning,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 9, pp. 1385-1392, Sept. 2006.

[18] T.S. Jaakkola and D. Haussler, “Exploiting Generative Models in Discriminative Classifiers,” *Proc. 1998 Conf. Advances in Neural Information Processing Systems II,* pp. 487-493, 1999.

[19] A.K. Jain and D.E. Zongker, “Feature Selection: Evaluation, Application, and Small Sample Performance,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 2, pp. 153-158, Feb. 1997.

[20] K. Kira and L.A. Rendell, “A Practical Approach to Feature Selection,” *Proc. Ninth Int’l Workshop Machine Learning (ICML ’92),* pp. 249-256, 1992.

[21] D. Koller and M. Sahami, “Toward Optimal Feature Selection,” *Proc. 13th Int’l Workshop Machine Learning (ICML ’96),* pp. 284-292, 1996.

[22] I. Kononenko, “Estimating Attributes: Analysis and Extensions of Relief,” *Proc. European Conf. Machine Learning (ECML ’94),* pp. 171-182, 1994.

[23] O.L. Mangasarian and D.R. Musicant, “Lagrangian Support Vector Machines,” *J. Machine Learning Research,* vol. 1, pp. 161-177, 2001.

[24] G. Qu, S. Hariri, and M.S. Yousif, “A New Dependency and Correlation Analysis for Features,” *IEEE Trans. Knowledge Data Eng.,* vol. 17, no. 9, pp. 1199-1207, Sept. 2005.

[25] M. Robnik-Sikonja and I. Kononenko, “Comprehensible Interpretation of Relief’s Estimates,” *Proc. 18th Int’l Conf. Machine Learning (ICML ’01),* pp. 433-440, 2001.

[26] M. Robnik-Sikonja and I. Kononenko, "Theoretical and Empirical Analysis of Relieff and Rrelieff," *Machine Learning,* vol. 53, nos. 1/2, pp. 23-69, 2003.

[27] G. Saon and M. Padmanabhan, "Minimum Bayes Error Feature Selection for Continuous Speech Recognition," *Proc. Advances in Neural Information Processing Systems 13 (NIPS '01),* pp. 800-806, 2001.

[28] Y. Sun, "Iterative Relief for Feature Weighting: Algorithms, Theories, and Applications," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 6, pp. 1035-1051, June 2007.

[29] K. Torkkola, "Feature Extraction by Non Parametric Mutual Information Maximization," *J. Machine Learning Research,* vol. 3, pp. 1415-1438, 2003.

[30] V.N. Vapnik, *Statistical Learning Theory.* Wiley-Interscience, Sept. 1998.

[31] N. Vasconcelos, "Feature Selection by Maximum Marginal Diversity: Optimality and Implications for Visual Recognition," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition (CVPR '03),* pp. 762-772, 2003.

[32] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the Zero Norm with Linear Models and Kernel Methods," *J. Machine Learning Research,* vol. 3, pp. 1439-1461, 2003.

[33] D. Wettschereck, D.W. Aha, and T. Mohri, "A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms," *Artificial Intelligence Rev.,* vol. 11, pp. 273-314, 1997.

[34] E.P. Xing, M.I. Jordan, and R.M. Karp, "Feature Selection for High-Dimensional Genomic Microarray Data," *Proc. 18th Int'l Conf. Machine Learning (ICML '01),* pp. 601-608, 2001.

[35] G. Xuan, X. Zhu, P. Chai, Z. Zhang, Y.Q. Shi, and D. Fu, "Feature Selection Based on the Bhattacharyya Distance," *Proc. 18th Int'l Conf. Pattern Recognition (ICPR '06),* pp. 1232-1235, 2006.

[36] S.-H. Yang and B.-G. Hu, "Feature Selection by Nonparametric Bayes Error Minimization," *Proc. 12th Pacific-Asian Conf. Knowledge Discovery and Data Mining (PAKDD '08),* pp. 417-428, 2008.

[37] S.-H. Yang and B.-G. Hu, "Efficient Feature Selection in the Presence of Outliers And Noises," *Proc. Asian Conf. Information Retrieval (AIRS '08),* pp. 188-195,

[38] S.-H. Yang, H. Zha, K.S. Zhou, and B.-G. Hu, "Variational Graph Embedding for Globally and Locally Consistent Feature Extraction," *Proc. European Conf. Machine Learning (ECML '09),* p. 538C553, 2009.

[39] S.-H. Yang and H. Zha, "Language Pyramid and Multi-scale Text Analysis," *Proc. 19th ACM Int'l Conf. Information and Knowledge Management (CIKM '10),* pp. 639-648, 2010.

[40] Y. Yang and X. Liu, "A Re-Examination of Text Categorization Methods," *Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '99),* pp. 42-49, 1999.

[41] Y. Yang and J.O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," *Proc. 14th Int'l Conf. Machine Learning (ICML '97),* pp. 412-420, 1997.

[42] H. Yoon, K. Yang, and C. Shahabi, "Feature Subset Selection and Feature Ranking for Multivariate Time Series," *IEEE Trans. Knowledge Data Eng.,* vol. 17, no. 9, pp. 1186-1198, Sept. 2005.

[43] K. Yu, X. Xu, M. Ester, and H.-P. Kriegel, "Feature Weighting and Instance Selection for Collaborative Filtering: An Information-Theoretic Approach∗," *Knowledge and Information Systems,* vol. 5, no. 2, pp. 201-224, 2003.

[44] L. Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," *J. Machine Learning Research,* vol. 5, pp. 1205-1224, 2004.

[45] Z. Zhao and H. Liu, "Spectral Feature Selection for Supervised and Unsupervised Learning," *Proc. 24th Int'l Conf. Machine Learning (ICML '07),* 2007.

Shuang-Hong Yang received the bachelor's degree from Wuhan University and the MS degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2005 and 2008, respectively. Currently, he is working toward the PhD degree at the College of Computing, Georgia Institute of Technology. His research interests include machine learning, data mining, and computational social science.

Bao-Gang Hu received the MSc degree from the Beijing University of Science and Technology, China, and the PhD degree from McMaster University, Canada, all in mechanical engineering, in 1983 and 1993, respectively. From 1994 to 1997, he was a research engineer and senior research engineer at C-CORE, Memorial University of Newfoundland, Canada. Currently, he is working as a professor with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Science, Beijing, China. His main research interests include intelligent systems, pattern recognition, and plant growth modeling. He is a senior member of the IEEE.