# Feature Weighting by RELIEF Based on Local Hyperplane Approximation

Hongmin Cai[1,⋆] and Michael Ng[2]

[1] South China University of Technology, Guangdong, P.R. China
caihongm@sysu.edu.cn
[2] Department of Mathematics, Hong Kong Baptist University, Hong Kong

**Abstract.** In this paper, we propose a new feature weighting algorithm through the classical RELIEF framework. The key idea is to estimate the feature weights through local approximation rather than global measurement, as used in previous methods. The weights obtained by our method are more robust to degradation of noisy features, even when the number of dimensions is huge. To demonstrate the performance of our method, we conduct experiments on classification by combining hyperplane KNN model (HKNN) and the proposed feature weight scheme. Empirical study on both synthetic and real-world data sets demonstrate the superior performance of the feature selection for supervised learning, and the effectiveness of our algorithm.

**Keywords:** Feature weighting, local hyperplane, RELIEF, Classification, KNN.

## 1 Introduction

Feature weighting plays an important step in the preprocessing of data, especially in data classification. In general, the feature weights are obtained by assigning a continuous relevance value to each feature via a learning algorithm by stressing on the context or domain knowledge. The feature weighting procedure is particularly useful for instance based learning models, which usually construct the distance metric by using all features. Moreover, feature weighting can reduce the risk of over-fitting by removing noisy features, thereby improve the predictive accuracy. Existing feature selection methods broadly falls into two categories, wrapper and filter methods. Wrapper methods use the predictive accuracy of predetermined classification algorithms (called base classifier), such as SVMs, as the criteria to determine the goodness of a subset of features [9,15]. Filter methods select features based on discriminant criteria that relies on the characteristics of data, independent of any classification algorithm [7,14,17]. The common discriminant criteria includes entropy measurement [18], Chi-squared

[⋆] This work was supported in part by NSFC under award number 60902076, NSF of Guangdong Province under award number 9451027501002551, and China Fundamental Research Funds for the Central Universities under award number $11lgpy33$.

P.-N. Tan et al. (Eds.): PAKDD 2012, Part II, LNAI 7302, pp. 335–346, 2012.

measurement [21], Fisher ratio measurement [10], mutual information measurement [20,4,3], and RELIEF-based measurement [19,27,28].

Due to the emerging needs in biomedical and bioinformatics areas, researchers are particularly interested in algorithms which can process of data with feature being of large (or huge) dimensions, such as, microarray scanning in cancer research. Therefore, filter methods are widely used due to its efficiency in computation. Among the existing filter methods in feature weighting, the RELIEF algorithm [19] is considered as one of the most successful ones due to its simplicity and effectiveness. The main idea behind RELIEF is to iteratively update feature weights by a distance margin to estimate the difference between neighboring patterns. It has been further generalized to average multiple, instead of just one, nearest neighbors when computing the sample margins, and was named as RELIEF-F [19]. The authors have shown that RELIEF-F can achieve significant improvement on performance of the original RELIEF [19]. Sun systematically proved that RELIEF is indeed an online algorithm for a convex optimization problem [27]. Through maximizing an averaged margin of nearest patterns in feature scaled space, RELIEF could estimate the feature weight in a straightforward and efficient manner. Based on the theoretical framework, one can impose outlier removal scheme called I-RELIEF since the margin averaging is sensitive to large variations [27]. To accomplish sparse feature weighting, the author introduced the $l_1$ penalty into optimization of I-RELIEF [28].

In this paper, we present a new feature weighting algorithm to extend classical RELIEF model. The main contribution of the proposed algorithm is that the feature weights are estimated from local patterns other than global ones, as used in exiting methods [19,27,28]. Therefore, the proposed feature weighting scheme is particularly useful when combined with local pattern based classifiers, such as HKNN [30], *ADAMENN* [8] and discriminant adaptive nearest neighbor (DANN) [16]. Besides, local patterns are more robust to the noises and outliers. It is promising to be used in applications where data are severely contaminated by noises or rich of redundance.

This paper is organized as follows. Section 2 introduces the background of the classical RELIEF method and its variations, including F-RELIEF and I-RELIEF. The main result is reported in this section. Section 3 demonstrates the performance of the proposed model. Extensive experiments have been conducted to compare with the classical methods on benchmark data sets. Conclusion is presented in Section 4.

## 2 The Proposed Method

### 2.1 RELIEF

The RELIEF algorithm has been successfully applied in feature weighing due to its simplicity and effectiveness [19,28]. The main idea of RELIEF is to iteratively adjust feature weights according to their ability to discriminate among neighboring patterns. Mathematically, suppose that $\boldsymbol{x}$ is a randomly selected sample of a binary class data. One can estimates its two nearest neighbors, wherein one is from its same class (called *the nearest hit* or NH) and the other is from a different class (called *the nearest miss* or NM). Then the weight $w_i$ for the $i$-th feature is updated by a heuristic estimation:

$$w_i = w_i + |x^{(i)} - NM^{(i)}| - |x^{(i)} - NH^{(i)}| \tag{1}$$

Since there is no exhaustive or iterative search evolved in RELIEF updating, this scheme is very efficient for the processing of data with huge dimensions, thus it is particularly promising for large-scale problems such as analysis of microarray data [24,28,7]. The authors have generalized the RELIEF model by averaging $k$, instead of just one in Eq. (1), nearest neighbors when computing the sample margins and was named as RELIEF-F model [19]. Experimental results have shown that RELIEF-F achieves superior performance over the original RELIEF. Its success is due to the robustness of margin estimation on multiple samples. However, the optimal number of nearest neighbors needs to be estimated empirically. Besides, RELIEF-F is also sensitive to noise degradation and the outliers. An benchmark achievement has been reported in [27], in which the author firstly proved that RELIEF is a convex optimization problem with a margin-based objective function,

$$\begin{aligned} &\max_{\boldsymbol{w}} \sum_{n=1}^{n} \rho_n(\boldsymbol{w}) \\ &\qquad := \textstyle\sum_{n=1}^{N}(\sum_{i=1}^{I} \omega_i |\boldsymbol{x}_n^{(i)} - NM^{(i)}(\boldsymbol{x}_n)| - \sum_{i=1}^{I} \omega_i |\boldsymbol{x}_n^{(i)} - NH^{(i)}(\boldsymbol{x}_n)|) \\ &\boldsymbol{s}.t. \quad \|\boldsymbol{w}\|_2^2 = 1, \qquad \boldsymbol{w} \geq 0 \end{aligned} \tag{2}$$

where $\rho_n = d(\boldsymbol{x}_n - NM(\boldsymbol{x}_n)) - d(\boldsymbol{x}_n - NH(\boldsymbol{x}_n))$ is defined as margin of a sample $\boldsymbol{x}_n$ for distance function $d(\boldsymbol{x}) = \sum_i |x_i|$. $NM(\boldsymbol{x}_n)$ and $NH(\boldsymbol{x}_n)$ are the nearest miss and hit for a sample $\boldsymbol{x}_n$, respectively.

To tackle the drawbacks of RELIEF, such as outlier detection and inaccurate updating, Sun reformulated the above problem as maximization of expected margin through scaling of features [27,28]:

$$\begin{aligned} \mathbf{E}[\rho(\boldsymbol{w})] &= \boldsymbol{w}^T(\underset{i \in NM}{\mathbf{E}}[|\boldsymbol{x}_n - \boldsymbol{x}_i|] - \underset{i \in NH}{\mathbf{E}}|[\boldsymbol{x}_n - \boldsymbol{x}_i|)] \\ &= \boldsymbol{w}^T \sum_{i \in NM} P(\boldsymbol{x}_i = NM(\boldsymbol{x}_n)|\boldsymbol{w})|\boldsymbol{x}_n - \boldsymbol{x}_i| - \sum_{i \in H_n} P(\boldsymbol{x}_i = NH(\boldsymbol{x}_n)|\boldsymbol{w})|\boldsymbol{x}_n - \boldsymbol{x}_i| \\ &= \boldsymbol{w}^T \boldsymbol{z}_n \end{aligned} \tag{3}$$

where $NM = \{i : 1 \leq i \leq N, y_i \neq y_n\}, NH = \{i : 1 \leq i \leq N, y_i = y_n, i = n\}$ are the sets of the nearest miss and the nearest hit, respectively. $P(\boldsymbol{x} = NM(\boldsymbol{x}_n)|W)$ (or $P(\boldsymbol{x} = NH(\boldsymbol{x}_n)|W)$) are the probabilities of the sample $\boldsymbol{x}$ being in the set of $NM(\boldsymbol{x}_n)$ (or $NH(\boldsymbol{x}_n)$) in the feature space scaled by weights $\boldsymbol{w}$. Though the probability distributions are unknown in prior, they can be estimated via kernel density estimation [6]. Empirical study has shown that the I-RELIEF achieves significant improvements over the traditional models. Task of classification on feature scaled dataset achieves higher accuracy than standard techniques such as SVM [12,15,9,26] and NN model [25]. Task of feature weighting is also robust to noisy features. In applications with a huge dimension of features, economic feature weights are appreciated not only because of computational consideration, but also most features being irrelevant [14,17]. To obtain sparse and economic feature weighting, the author introduced the $l_1$ penalty into the optimization of I-RELIEF [28].

However, since the expectation in Eq. (3) is carried out on the set of nearest miss or hit, which consisted of the nearest neighbors of all observed samples, the feature weight estimation may be less inaccurate if the samples contain many outliers, or most of the features are being irrelevant. In both cases, the distance between the tested one and its nearest neighbors are in large value. It follows that large bias will be introduced in margin estimation via averaging operation. Although one can reduce the influence of the abnormal samples by introducing kernel distribution estimation [27,28], it will introduce additional free parameter estimation. Moreover, probability estimation via kernel approximation is sensitive to the sample size [6,13]. Therefore, it limits the empirical applications such as in analysis of microarray data, in which the data is notoriously known for that the dimension of sample observation is far less than that of the sample feature [11]. In this paper, we propose to use a local hyperplane to approximate the set of the nearest hit and miss and then estimate the feature weight through maximization of an expected margin defined by the hyperplane. The contribution of this approximation is that the hyperplane is more robust for noisy features degradation than averaging over all neighbors [19,27,28].

### 2.2 Approximation by Local Hyperplane

Given a sample $\boldsymbol{x}$, it can be represented by a local hyperplane of class $c$ by:

$$LH_c(x) = \{\boldsymbol{s} \mid \boldsymbol{s} = \boldsymbol{H\alpha}\}, \tag{4}$$

where $\boldsymbol{H}$ is a $I \times n$ matrix composed by $n$ NNs of the sample $\boldsymbol{x}$: $\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n\}$, with $\boldsymbol{h}_i$ being the $i$-th nearest neighbor (called *prototype*) of class $c$. The parameter of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ is the weights of the prototypes $\{\boldsymbol{h}_i,\ i = 1, 2, \ldots, n\}$. It can be viewed as spanning coefficients of the subspace $LH_c(\boldsymbol{x})$. Therefore, the hyperplane can be represented as: $\{\cdot \mid \boldsymbol{H\alpha} = \alpha_1\boldsymbol{h}_1 + \alpha_2\boldsymbol{h}_2 + \ldots + \alpha_n\boldsymbol{h}_n\}$. The value of $\boldsymbol{\alpha}$ is solved by minimizing the distance between the sample $\boldsymbol{x}$ and its local hyperplane of $LH_c(\boldsymbol{x})$ within feature scaled space:

$$\begin{aligned} &J_c(\boldsymbol{\alpha}) = arg\min \frac{1}{2}\sum_{i=1}^{I} \omega(i)(x_i - s_i)^2 = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{H\alpha})^T \boldsymbol{W}(\boldsymbol{x} - \boldsymbol{H\alpha}) \\ &\quad Subject\ to: \\ &\sum_{i=1}^{k} \alpha_i = 1\ ,\quad \boldsymbol{\alpha} \geq 0 \end{aligned} \tag{5}$$

where $\boldsymbol{s} = (s_1, s_2, \ldots, s_I) = \boldsymbol{H\alpha} \in LH_c(\boldsymbol{x})$. $\boldsymbol{W}$ is a diagonal matrix with diagonal elements $w_i$ being the weight of the $i$-th feature.

We are proposing to use the hyper plane to represent the set of nearest miss $NM(\boldsymbol{x})$ and nearest hit $NH(\boldsymbol{x})$ for the given sample $\boldsymbol{x}$. The beneficiary of the representation is to characterize the local sample patterns robustly. Then the distance between the sample to its $NH$ (or $NM$) set can be estimated from its local hyperplane other than averaging across over all samples within the set. Therefore, we redefine the margin for a sample $\boldsymbol{x}$ as $\rho_n = d(\boldsymbol{x}_n - LH_{NM}(\boldsymbol{x}_n)) - d(\boldsymbol{x}_n - LH_{NH}(\boldsymbol{x}_n))$. The feature weights are now estimated through maximization of total margins:

$$\begin{aligned}
\max_{\boldsymbol{w}} \mathbf{E}[\rho(\boldsymbol{w})] &= \frac{1}{N} \max_{\boldsymbol{w}} \sum_{n=1}^{N} (\sum_{i=1}^{I} \omega_i |\boldsymbol{x}_n^{(i)} - LH_{NM}^{(i)}(\boldsymbol{x}_n)| - \sum_{i=1}^{I} \omega_i |\boldsymbol{x}_n^{(i)} - LH_{NH}^{(i)}(\boldsymbol{x}_n)|) \\
&= \boldsymbol{w}^T \frac{1}{N} \max_{\boldsymbol{w}} \sum_{n=1}^{N} (\sum_{i=1}^{I} |\boldsymbol{x}_n^{(i)} - \boldsymbol{\alpha} \boldsymbol{H}_{NM}^{(i)}(\boldsymbol{x}_n)| - \sum_{i=1}^{I} |\boldsymbol{x}_n^{(i)} - \boldsymbol{\beta} \boldsymbol{H}_{NH}^{(i)}(\boldsymbol{x}_n)|) \\
&= \boldsymbol{w}^T \boldsymbol{z}_n
\end{aligned} \tag{6}$$

where $\boldsymbol{H}_{NM}(\boldsymbol{x}_n)$ and $\boldsymbol{H}_{NH}(\boldsymbol{x}_n)$ are the nearest neighbors for the set of the nearest miss and hit of the sample $\boldsymbol{x}_n$. $\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}_n$ are the coefficients for spanning of hyperplane $LH_{NM}^{(n)}$ and $LH_{NH}^{(n)}$. $\boldsymbol{w}$ is a vector with its $i$-th element $\boldsymbol{w}(i)$ being the weight of the $i$-th feature, for $i = 1, 2, \ldots, I$. To solve the minimization problem of Eq. (6), one should estimate the parameters of $\boldsymbol{\alpha}_n$, $\boldsymbol{\beta}_n$, which are dependent on the nearest neighborhoods. The main problem of the estimation, however, is that the nearest neighbors of a given sample are unknown before learning. In the presence of many thousands of irrelevant features, the nearest neighbors defined in the original space can be completely different from those in the induced space. Therefore, the nearest neighbors defined in the original feature space may not be true in the weighted feature space. To solve the difficulties, we have designed an iterative algorithm, similar to the EM algorithm and I-RELIEF [27], to achieve the goal.

**Step 1:** In $t$-th iteration, for a given sample $\boldsymbol{x}$, we estimate the parameter of $\boldsymbol{\alpha}$ by constructing the local hyperplane of the nearest hit set within induced feature space. It is trivial to show that the minimization of Eq. (5) is equivalent to solving the following quadratic programming:

$$\begin{aligned}
&\min_{\boldsymbol{\alpha}} \frac{1}{2} \boldsymbol{\alpha}^T \bar{\boldsymbol{H}} \boldsymbol{\alpha} + \boldsymbol{f}^T \boldsymbol{\alpha} \\
&s.t. \quad \mathbf{1}^T \boldsymbol{\alpha} = 1, \; \boldsymbol{\alpha} \geq 0
\end{aligned} \tag{7}$$

where $\bar{\boldsymbol{H}} = \boldsymbol{H}^T \boldsymbol{W}^{(i)} \boldsymbol{H}$, $\boldsymbol{f} = -\boldsymbol{x}^T \boldsymbol{W}^{(i)} \boldsymbol{H}$, and $\mathbf{1}$ is an unitary vector whose elements are all being 1. The matrix of $W^{(i)}$ is the $t$-th feature weight matrix, satisfying $W^{(i)} \mathbf{1} = \boldsymbol{w}$. The parameter of $\boldsymbol{\beta}$ for nearest miss hyperplane is obtained similarly. Minimization of Eq. (5) is a constrained quadratic program problem and standard techniques can be used to obtain its solution. In particular, since the matrix of $\bar{\boldsymbol{H}}$ is symmetric and non-negative, the minimization could be solved efficiently through standard techniques, such as active set [23].

**Step 2:** Estimation of the total margin with respect to $\boldsymbol{w}^{(i)}$.

$$\rho(\boldsymbol{w}^{(i)}) = \frac{1}{N}\sum_{n=1}^{N}(\sum_{i=1}^{I}\omega_i|\boldsymbol{x}_n^{(i)} - \boldsymbol{\alpha}\boldsymbol{H}_{NM}^{(i)}(\boldsymbol{x}_n)| - \sum_{i=1}^{I}\omega_i|\boldsymbol{x}_n^{(i)} - \boldsymbol{\beta}\boldsymbol{H}_{NH}^{(i)}(\boldsymbol{x}_n)|) \quad (8)$$

**Step 3:** Estimation of the weight $\boldsymbol{W}$ in $(i+1)$-th iteration.

$$\begin{aligned}\boldsymbol{w} &= arg\max_{\boldsymbol{w}}\rho(\boldsymbol{w}^{(i)})\\ &= \frac{1}{N}\sum_{n=1}^{N}(\sum_{i=1}^{I}\omega_i|\boldsymbol{x}_n^{(i)} - \boldsymbol{\alpha}\boldsymbol{H}_{NM}^{(i)}(\boldsymbol{x}_n)| - \sum_{i=1}^{I}\omega_i|\boldsymbol{x}_n^{(i)} - \boldsymbol{\beta}\boldsymbol{H}_{NH}^{(i)}(\boldsymbol{x}_n)|) \quad (9)\end{aligned}$$

The above steps iterate alternatively until their convergence. The last two steps are similar to the one used in I-RELIEF [27], and we name our scheme as **LH-RELIEF** since it requires a local hyperplane approximation.

The pseudo-code for the LH-RLIEF is summarized in Alg. (2.1)

**Algorithm 2.1:** LH-RELIEF ALGORITHM$(V, W, \lambda)$

```
comment: Variables Initialization: $w=\frac{1}{I}$, stopping criteria$\epsilon$and number of iterations $T$
for $t \leftarrow 1$ to $T$
while $\|w^{(t+1)} - w^{(t)}\| > \epsilon$
     { 1. Estimate the coefficients for hyperplane of nearest miss and hit $\alpha, \beta$
  do { 2. Calculate the margin by Eq .(8)
     { 3. Update the weights by Eq. (9)
return $(w)$
```

## 3 Experimental Results

We shall demonstrate the performance of the proposed scheme through classification evaluation on both synthetic and empirical problems. In particular, we are interested in its: 1) performance of classification compared with other feature weighting scheme; 2) robustness when processing the samples with irrelevant features of large dimension.

### 3.1 Selection of Classifier

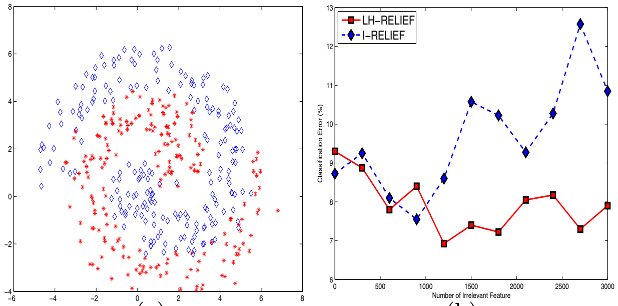
In our experiments, we selected the hierarchical $k$-nearest neighbor (HKNN) algorithm to conduct the comparison on feature weighting [30]. HKNN could be viewed as a localized approximation of $K$-nearest neighbor model. In this model, each class is modeled as a smooth and low-dimensional manifold embedded in the high-dimensional data space by assuming that the manifolds are locally linear.

There are two steps involved in classification by HKNN. In the first step, for each tested sample, it constructs local hyperplanes for each class. The label of the tested sample is assigned to the class whose local hyperplane to the tested sample is minimized. Empirical study has shown that the HKNN produced a comparable or even better performance of classification than standard techniques, including KNN and SVM [30,8,29]. One may note that the HKNN model shares the similar idea with our approach in that the sample information is inferred from local structure, which is the main reason for us to choose this particular classifier.

Since the HKNN model does not consider the influence of feature weights, the test data will be firstly scaled into feature space before the classification is carried out. The hyper-parameters used in training phase are estimated through ten-fold cross validation.

### 3.2 Fermat's Spiral Problem

In the first example, we shall test the performance of the proposed method on the well-known Fermat's Spiral problem. The test dataset consists of two classes with 200 samples for each class. The labels of the Spiral are completely determined by its first two features. The shape of the Fermat's Spiral distribution is shown in Fig. 1(a). Heuristically, the label of a sample will be inferred easily from its local neighbors. Classification based on local information will give more accurate assignment than global measurement based prediction (or classification) does since the later one is sensitive to noise degradation. To tackle this drawback, Sun proposed to lower the influence of the samples nearby through modeling of their probability distribution via kernel techniques [27]. This strategy is straightforward and successful. However, if the dominant (informative) features are buried by the irrelevant (less informative) ones, estimation of the probability via distance will be less accurate since the irrelevant feature may introduce a large variation to distance, for instance, the irrelevant features are being in a huge dimension. In order to show this, irrelevant features following standard norm distribution are added to the Spiral for classification testing. The dimensions of irrelevant features are ranging from $\{0, 300, 600, 900, 1200, 1500, 1800, 2100, 2400, 2700, 3000\}$. Two feature weighting scheme, I-RELIEF and LH-RELIEF were firstly applied to quantify the importance of feature. Then the classification was performed on dataset scaled by the feature weights. For each experiment, ten folds cross validation scheme is used to compute the accuracy of classification. To eliminate the statistical variations, we have conducted ten times experiments independently on each dataset and averaged classification error is recorded and, shown in Fig. 1(b). We observe that, the performance of the two methods are very similar when the dimension of the irrelevant features is small. However, if the dimension of irrelevant features tends to be large, the performance of I-RELIEF is severely degraded by the noises. In comparison, the performance of LH-RELIEF is very stable and produces superior outcomes.



(a) (b)

**Fig. 1.** Experiment on Fermat's Spiral. (a) Distribution of binary Fermat's Spiral problem. Each class has 200 samples and is labeled by different colors; (b) Irrelevant features with verified dimensions are added to test the robustness of the feature weighting schemes. The LH-RELIEF outperforms I-RELIEF with respect to classification error. With the increase of dimension of the irrelevant features, the performance of I-RELIEF is degraded while LH-RELIEF keeps stable.

### 3.3 UCI Data Sets

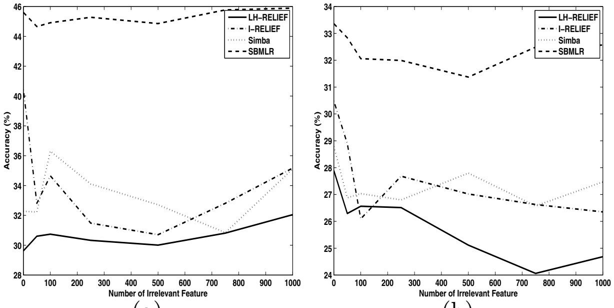
In the second experiment, we tested the proposed technique on ten medium sized datasets. The tested benchmark data sets were downloaded from the UCI Machine Learning Repository [1], and they have been widely tested by various classification benchmark models. The characteristics of the datasets are summarized in Table 1. We compare our algorithm with four other algorithms, including Iterative Search Margin Based Algorithm (Simba) [2], sparse Bayesian multinomial logistic regression (SBMLR) [5] and I-RELIEF [27]. Simba is a local learning based algorithm similar to RELIEF. SBMLR is a special kind of sparse multinomial logistic regression models with Bayesian regularization. Multinomial logistic regression algorithm has been successfully used in text processing [31] and microarray classification [22]. The beneficiary of adding regularization parameter into sparse multinomial logistic regression via a Laplace prior is that an analytical solution could be obtained. Besides, its performance is similar to using cross-validation based model selection, thus greatly reducing computational expense.

For each dataset, the optimal parameters were estimated by ten-fold cross validation. The obtained feature weights under optimal parameters were used to scale the raw datasets. Twenty times experiments on each dataset were performed independently and classification errors were averaged to evaluate the performance of the feature weighting scheme. We will use the classification error to quantify the discrimination power of weighting scheme. Furthermore, statistical testing is also useful to fully comprise the performance of feature weights [27]. We selected the Students paired two-tailed $t$-test to achieve the goal. The $p$-value of the $t$-test represents the probability that two sets of compared results come from distributions with an equal mean. In this experiment, a $p$-value of 0.05 is considered statistically significant.

The results are summarized in Table 2. We observe that that LH-RELIEF and I-RELIEF are statistically different from the tested ten datasets. The performance of classification after LH-RELIEF is better than after I-RELIEF in 9 of 10 experiments. Among the four feature weighting schemes, LH-RELIEF outperforms others in 5 of 10 datasets, while almost is suboptimal in other five dataset.

**Table 1.** Summary of tested datasets and their characteristics

| Data set | #Instances | #Classes | #Feature |
|---|---|---|---|
| Bupa | 345 | 2 | 6 |
| Teach | 151 | 3 | 5 |
| Sonar | 208 | 2 | 60 |
| Cancer | 198 | 6 | 32 |
| Prokaryotic | 997 | 3 | 20 |
| Eukaryotic | 2427 | 4 | 20 |
| Haberman | 306 | 2 | 3 |
| Page block | 5473 | 5 | 10 |
| Pima | 768 | 2 | 8 |
| Spambase | 4601 | 2 | 57 |



(a) (b)

**Fig. 2.** Experiment on benchmark dataset of Bupa and Pima by adding irrelevant features in verified dimension, extending from 0 to 1000. (a) Bupa; (b) Pima.

In the last experiment, we are willing to test the performance of the algorithm on data in huge dimensions. More specifically, we are interested in the robustness of the algorithm on feature weighting with respect to the dimension of the irrelevant features. We selected two test datasets: Bupa and Pima. For each dataset, irrelevant features are added to the raw dataset. The added irrelevant features are independently sampled from zero-mean and unit-variance Gaussian distribution. Their dimensions are ranged from 0 to 1000. Including useless features is

**Table 2.** Classification accuracies (%) on 10 real data sets. The LH-RELIEF shows to be statistically different from the I-RELIEF in 9 among 10 datasets. The $P$-value for each dataset is shown in parenthesis. Overall, the better results are subscripted by star under different feature weighting scheme. The LH-Relief outperforms the standard ones in most cases when the two methods show a statistically difference.

| Dataset | LH-RELIEF | I-RELIEF<br>($P$-value) | SBMLR | Simba |
|---|---|---|---|---|
| Bupa | 69.7* | 66.7 (0.00) | 56.2 | 66.8 |
| Teach | 64.4* | 46.3 (0.00) | 34.4 | 62.3 |
| Sonar | 86.7* | 84.3 (0.00) | 82.7 | 85.7 |
| Cancer | 76.2 | 76.0 (0.48) | 76.9* | 76.4 |
| Prokaryotic | 90.5* | 89.8 (0.00) | 90.4 | 89.3 |
| Eukaryotic | 82.8 | 81.2 (0.00) | 83.5* | 81.3 |
| Haberman | 69.3 | 72.3 (0.00) | 69.9 | 68.7 |
| Page Block | 94.5 | 94.1 (0.00) | 95.7* | 89.8 |
| Pima | 74.0 | 70.3 (0.00) | 68.9 | 74.5* |
| Spambase | 84.8* | 78.0 (0.00) | 79.3 | 39.4 |

less appreciated in applications where the acquisition of data is quite expensive. For example, it may complicate the pathway research if irrelevant genes are included in microarray data analysis [27]. We would welcome such complication in order to show the robustness of the algorithm.

The hyper-parameters, such as the kernel size $\sigma$ in I-RELIEF and the number of nearest neighbors $k$ in LH-RELIEF are estimated through ten-fold cross validation. To eliminate statistical variations, each algorithm is run for twenty times on each noisy dataset. In each run, a dataset is randomly partitioned into training and testing. The averaged testing errors serve as the criterion to quantify the performance of the algorithm, and the results are drawn in Fig.2. For Bupa, the classification error of the classifier after LH-RELIEF is smaller than that after I-RELIEF in all dimensions, Fig. 2(a). This observation is coincided with the results in Table. 2, implying that the feature weights estimated by LH-RELIEF are more accurate and robust to the noises. For Pima, the performance of the two scheme is almost comparable when the dimension of the the irrelevant features is small, Fig. 2(b). However, the testing error after LH-RELIEF dramatically decreased with respect to the dimension of the irreverent features. In comparison, the classification error after I-RELIEF tends to be greater. The experiment further demonstrates that the proposed feature weighting scheme is more immune to the noisy features by showing surprising high degree of robustness.

## 4 Discussion

In this paper, we proposed a new feature weight scheme to tackle the common drawbacks of the RELIEF family. The nearest miss and hit subset are approximated by constructing a local hyperplane. Then the updating of feature weights is achieved by measuring the margin between the sample and its hyperplane under general RELIEF framework. The main contribution of the new variation is that the margin is more robust to the noises and the outliers than earlier works do. Therefore, the feature weights can characterize the local structure more accurately. Experimental results on both synthetic and real-world datasets validate our findings. The proposed weighting scheme performs superior on most test data with respect to classification error. We also observed that the algorithm was convergent in most cases, though theoretical justification is needed.

## References

1. Asuncion, A., Newman, D.: UCI machine learning repository (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
2. Bachrach, G.R., Navot, A., Tishby, N.: Margin Based Feature Selection - Theory and Algorithms. In: Proc. 21st International Conference on Machine Learning (ICML), pp. 43–50 (2004)
3. Brown, G.: An Information Theoretic Perspective on Multiple Classifier Systems. In: Benediktsson, J.A., Kittler, J., Roli, F. (eds.) MCS 2009. LNCS, vol. 5519, pp. 344–353. Springer, Heidelberg (2009)
4. Brown, G.: Some Thoughts at the Interface of Ensemble Methods and Feature Selection. In: El Gayar, N., Kittler, J., Roli, F. (eds.) MCS 2010. LNCS, vol. 5997, pp. 314–314. Springer, Heidelberg (2010)
5. Cawley, G.C., Talbot, N.L.C., Girolami, M.: Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation. Advances in Neural Information Processing Systems 19 (2007)
6. Christopher, A., Andrew, M., Stefan, S.: Locally weighted learning. Artificial Intelligence Review 11, 11–73 (1997)
7. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology 3(2), 185–205 (2005)
8. Domeniconi, C., Peng, J., Gunopulos, D.: Locally adaptive metric nearest-neighbor classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(9), 1281–1285 (2002)
9. Duan, K.B.B., Rajapakse, J.C., Wang, H., Azuaje, F.: Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Transactions on Nanobioscience 4(3), 228–234 (2005)
10. Duda, R., Hart, P., Stork, D.: Pattern Classification. Wiley (2001)
11. Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97(458), 611–631 (2002)
12. Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. BMC bioinformatics 16, 906–914 (2000)
13. Girolami, M., He, C.: Probability density estimation from optimally condensed data samples. IEEE Transactions on Pattern Analysis and Machine Intelligence 25, 1253–1264 (2003)
14. Guyon, I.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003)

15. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning 46, 389–422 (2002)
16. Hastie, T., Tibshirani, R.: Discriminant adaptive nearest neighbor classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 18, 607–616 (1996)
17. Huang, C.J., Yang, D.X., Chuang, Y.T.: Application of wrapper approach and composite classifier to the stock trend prediction. Expert Systems with Applications 34(4), 2870–2878 (2008)
18. Koller, D., Sahami, M.: Toward optimal feature selection. In: Saitta, L. (ed.) Proceedings of the Thirteenth International Conference on Machine Learning (ICML), pp. 284–292. Morgan Kaufmann Publishers (1996)
19. Kononenko, I.: Estimating Attributes: Analysis and Extensions of RELIEF. In: Bergadano, F., De Raedt, L. (eds.) ECML 1994. LNCS, vol. 784, pp. 171–182. Springer, Heidelberg (1994)
20. Kwak, N., Choi, C.H.: Input feature selection by mutual information based on parzen window. IEEE Trans. Pattern Anal. Mach. Intell. 24, 1667–1671 (2002)
21. Liu, H., Setiono, R.: Feature selection via discretization. IEEE Transactions on Knowledge and Data Engineering 9, 642–645 (1997)
22. Narlikar, L., Hartemink, A.J.: Sequence features of dna binding sites reveal structural class of associated transcription factor. Bioinformatics 22(2), 157–163 (2006)
23. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer (August 2000)
24. Peng, Y.H.: A novel ensemble machine learning for robust microarray data classification. Computers in Biology and Medicine 36, 553–573 (2006)
25. Shakhnarovich, G., Darrell, T., Indyk, P. (eds.): Nearest-Neighbor Methods in Learning and Vision: Theory and Practice. MIT Press (2006)
26. Statnikov, A., Wang, L., Aliferis, C.F.: A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9, 319–328 (2008)
27. Sun, Y.: Iterative relief for feature weighting: Algorithms, theories, and applications. IEEE Trans. Pattern Anal. Mach. Intell. 29(6), 1035–1051 (2007)
28. Sun, Y., Todorovic, S., Goodison, S.: Local-learning-based feature selection for high-dimensional data analysis. IEEE Trans. Pattern Anal. Mach. Intell. 32(9), 1610–1626 (2010)
29. Tao, Y., Vojislav, K.: Adaptive local hyperplane classification. Neurocomputing 71(13-15), 3001–3004 (2008)
30. Vincent, P., Bengio, Y.: K-local hyperplane and convex distance nearest neighbor algorithms. In: Advances in Neural Information Processing Systems, pp. 985–992. The MIT Press (2001)
31. Zhang, T., Oles, F.J.: Text categorization based on regularized linear classification methods. Information Retrieval 4(1), 5–31 (2001)