Journal of Applied Sciences

ISSN 1812-5654

Journal of Applied Sciences 9 (6): 1014-1024, 2009
ISSN 1812-5654

Feature Ranking by Weighting and ISE Criterion of Nonparametric Density Estimation

[1]Xiaoming Wang and [1,2]Shitong Wang
[1]School of Information Engineering, Southern Yangtze University, Wuxi, China
[2]National Key Laboratory of CAD and CG, ZheJiang University, HangZhou, China

Abstract: This study deals with how to efficiently rank features of datasets. As we may know well, reducing the dimensionality of datasets (i.e., feature reduction) is an important step in pattern recognition tasks and exploratory data analysis. Quite often, feature ranking is required before completing feature reduction. In this study, a novel classifier-free feature ranking approach based on the combination of both weighting features and ISE (Integrated Squared Error) criterion is proposed. ISE is measured in terms of the modified non-parametric Parzen window density estimator in this study. The advantage of the proposed approach is that it allows us to make an efficient and effective non-parametric implementation and requires no prior assumption. The experimental results demonstrate that the proposed approach here is very promising.

Key words: Dimensionality reduction, ISE criterion, non-parametric Parzen window density estimation, unsupervised learning

INTRODUCTION

Dimensionality reduction is essential to improve the accuracy, efficiency and scalability of a classification process in pattern recognition tasks and exploratory data analysis. In order to realize an effective dimensionality reduction, feature ranking is quite often required. There are two categories of feature ranking strategies: The wrapper strategy and the filter strategy (Wang *et al*., 1999; Guyon and Elisseeff, 2003; Song *et al*., 2005). In the wrapper strategy, feature ranking associated with a particular classifier can be done by evaluating the classification accuracy. The huge computation burden is the major drawback of this strategy. The filter strategy is to rank features by evaluating some criterion. Such a criterion may be the classifier-parameter-based criterion (Basak *et al*., 1998; Wang *et al*., 1999; Andonie and Cataron, 2005) and the classifier-free criterion (Dash *et al*., 1997; Morita *et al*., 2003; Torkkola, 2003; Dy and Bradley, 2004; Biesiada *et al*., 2005). The latter is our concern in this study.

In general, most of the classifier-free criteria (Morita *et al*., 2003; Torkkola, 2003; Dy and Bradley, 2004; Biesiada *et al*., 2005) are based on the first-order or second-order statistics computed from the empirical distributions. However, these criteria based on the first-order statistics are sensitive to noise in datasets, whereas, these criteria based on the second-order statistics are sensitive to data transformation. In order to circumvent these drawbacks and enhance the robustness to noise and data transformation, MI (mutual information) as the high order statistics has been introduced into the classifier-free criteria. However, this poses great difficulties as it requires prior knowledge of the underlying probability density functions of datasets and the numerical integration of these functions, which leads to a high computational complexity. Moreover, MI-based criteria often fail in higher dimensions (Kononenko, 1994; Torkkola, 2003; Andonie and Cataron, 2005). In the latest advance of MI-based criteria, Torkkola (2003) proposed the non-parametric MI criterion based on Parzen-window density estimation (Bishop, 1995) and applied it to feature transformation. Similar studies are also described by Principe *et al*. (2000), Guyon and Elisseeff (2003) and Huang and Chow (2003).

As pointed out, this study deals with feature ranking (as the first step of feature extraction). In this study, a novel feature ranking approach is proposed. The core of the proposed approach here can be stated in brief as follows. First, a weight with one as its initial value is assigned to every feature. The weight of each feature reflects its importance in the data space. Thus, with the help of the concept of feature transform (Torkkola, 2003), two data spaces are generated. One is associated with all the original features and the other is associated with all the weighted features. Second, we use the Parzen window density estimator to derive two underlying probability density functions of the datasets, respectively, in these two data spaces. In order to keep more information in the datasets, we adopt the more precise Parzen window

Corresponding Author: Xiaoming Wang, School of Information Engineering, Southern Yangtze University, Wuxi, China
Tel: 86-510-85912151 Fax: 86-510-85912136

density estimator generated by using the fuzzy C-means algorithm (FCM) (Dunn, 1973; Bezdek, 1981) and then computing the variances of the corresponding different clusters (i.e., classes). Algorithm FCM realizes clustering by minimizing its objective function which induces the corresponding fuzzy membership functions. It was first developed by Dunn (1973) and then improved by Bezdek (1981). It has been frequently used in pattern recognition, data mining and so on. Third, we use ISE (Integrated Squared Error) criterion to measure the global error between the density estimates of these two data spaces. ISE is a robust criterion for the presence of noise and outliers in datasets (Girolami and Chao, 2003; Wang *et al*., 2008). Finally, by using the gradient descent method for ISE, we derive the iterative learning rules for the used weights and final weights will also be accordingly obtained.

The proposed approach seems to be very much related with Torkkola (2003) study. However, the distinctive characteristics of the proposed approach here should be emphasized: First, the proposed approach is the classifier-free one, i.e., no prior knowledge about the class labels in datasets is assumed, whereas, Torkkola's approach is oriented to such datasets where the class labels of all the data in datasets are known. This indicates the big discrepancy between the proposed approach and his study. The used MI criterion in his study is only related to the low-dimensional weighted data space, whereas, the used ISE in the proposed approach deals with the global error measure between the original data space and the weighted data space with the same dimension. Moreover, from the pure viewpoint of the computational complexity, ISE may have less computational burden than MI criterion, we will explain the reason later. Second, the proposed approach adapts FCM to partition the dataset to be analysed and then calculate the variance of every cluster obtained. Thus, the underlying probability density function computed by the modified Parzen window density estimator will make use of more information in datasets.

THE PARZEN WINDOW DENSITY ESTIMATOR AND ISE CRITERION

The modified parzen window density estimator in the proposed approach: The traditional Parzen window density estimator is in particular attractive when no a prior information is available to guide the choice of the precise form of the density of a dataset to be analysed. A probability density estimate $\hat{p}_N(x)$ can be obtained from the finite d-dimensional data points $(x_1,x_2,...,x_N \in X^d)$ by employing the following Parzen window density estimator:

$$\hat{p}_N(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\phi(\frac{x-x_i}{h_N}) \tag{1}$$

where, $x_i \in X^d$, N is the number of all the data points in the dataset, $\phi(\cdot)$ is a kernel window function and h_N is the window's width and V_N is the window's volume.

In this study,

$$\frac{1}{V_N}\phi(\frac{x-x_i}{h_N})$$

in Eq. 1 is taken as the following Gaussian kernel function:

$$G(x,\Sigma^{'}) = \frac{1}{(2\pi)^{\frac{d}{2}}h^d\,|\Sigma^{'}|^{\frac{1}{2}}}\exp(-\frac{x^T\Sigma^{'-1}x}{2h^2}) \tag{2}$$

where, Σ' is the covariance matrix, h is the window's width. Let $\Sigma = h^2\Sigma'$, since

$$\frac{(\Sigma')^{-1}}{h^2} = (h^2\Sigma^{'})^{-1}$$

so, we have:

$$G(x,\Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}\,|\Sigma|^{\frac{1}{2}}}\exp(-x^T\Sigma^{-1}x) \tag{3}$$

Thus, we have

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}G(x-x_i,\Sigma) \tag{4}$$

In most cases, in order to ease the computational complexity, one adopts the unified variance, i.e., $\Sigma = \sigma^2 I$ where, I is the identity matrix, σ^2 is the variance of the dataset. However, when the dataset contains several classes where the compactness of every class is different, which actually implies that the variance of every class is different from others, this assumption may pose a serious issue, i.e., the obtained density estimate loses much useful information contained in the dataset.

In particular, due to high dimensionality, all the data points in a dataset are more sparsely distributed in a high dimensional data space such that the compactness of all the projected data points along every dimension may perhaps be different from others. Therefore, we should compute Σ in Eq. 4 based on the variance of every dimension.

Next, let us state the modified Parzen window density estimator used in the proposed approach.

Given the dataset $D_X = \{x_1,...,x_N\}$, $x_i = (x_{i1}, x_{i2},...,x_{id})^T \in X^d$, i = 1,2,..., N, suppose there are N_c classes in D_x and there are J_c data points in the cth class, then we can estimate the probability P(c) of the cth class as:

$$P(c) = \frac{J_c}{N} \qquad 1 \le c \le N_c \tag{5}$$

We employ Eq. 4 to estimate the conditional probability function of the cth class as:

$$p(x \mid c) = \frac{1}{J_c}\sum_{i=1}^{J_c} G(x - x_{ci}, S_c) \tag{6}$$

where, x_{ci} is the value of the ith data point in the cth class; Σ_c denotes the diagonal covariance matrix of the cth class and it is defined as:

$$\Sigma_c = h^2 * \begin{bmatrix} \sigma_{c1}^2 & & \\ & \ddots & \\ & & \sigma_{cd}^2 \end{bmatrix} \tag{7}$$

where, h is the window's width, * denotes the multiplication operator, σ_{ck}^2, $k = 1,2,\cdots d,$ is the variance of all the data points in the cth class along the kth dimension, which is approximately calculated using:

$$\sigma_{ck}^2 = \frac{1}{J_c - 1}\sum_{i=1}^{J_c}(x_{cik} - \overline{x}_{ck})^2 \tag{8}$$

$$\overline{x}_{ck} = \frac{1}{J_c}\sum_{i=1}^{J_c} x_{cik} \tag{9}$$

Thus, the density function p(x) for the data D_X can be estimated as:

$$\begin{aligned} p(x) &= \sum_{c=1}^{N_c} P(c)p(x \mid c) \\ &= \sum_{c=1}^{N_c}\left(\frac{J_c}{N} * \frac{1}{J_c}\sum_{i=1}^{J_c} G(x - x_{ci}, S_c)\right) \\ &= \frac{1}{N}\sum_{c=1}^{N_c}\sum_{i=1}^{J_c} G(x - x_{ci}, S_c) \end{aligned} \tag{10}$$

ISE criterion in the proposed approach: In order to evaluate the importance of every dimension (feature), we utilize the following weighting strategy, i.e., imposing a linear dimension-keeping transformation $y = w \otimes x = (w_1x_1, w_2x_2, ..., w_dx_d)^T$ to x, where, $w = (w_1, w_2, ..., w_d)^T$,

$$\sum_{i=1}^{d} w_i = d,$$

$w_i > 0$, i = 1,2,...,d. The bigger w_i is, the more important the corresponding ith feature will be. With this linear transformation, the original dataset D_X becomes a transformed dataset D_Y.

For this transformed dataset D_Y, we can estimate its probability density function q(x) using Eq. 10 as:

$$q(x) = \frac{1}{N}\sum_{c=1}^{N_c}\sum_{i=1}^{J_c} G(x - y_{ci}, \Sigma_c') \tag{11}$$

where, S_c' takes a diagonal covariance matrix for ease of the computational complexity, which is defined as:

$$\Sigma_c' = h^2 * \begin{bmatrix} \sigma_{c1}'^2 & & \\ & \ddots & \\ & & \sigma_{cd}'^2 \end{bmatrix} \tag{12}$$

where, h is the window's width, $\sigma_{ck}'^2$ $(k = 1,2,\cdots,d)$ is the variance of all the data points in the cth class along the kth transformed dimension, which can be estimated as:

$$\sigma_{ck}'^2 = \frac{1}{J_c - 1}\sum_{i=1}^{J_c}(y_{cik} - \overline{y}_{ck})^2 = w_k^2\sigma_{ck}^2 \tag{13}$$

where, $\overline{y}_{ck} = \frac{1}{J_c}\sum_{i=1}^{J_c} w_k x_{cik}$ and $y_{cik} = w_k x_{cik}$. That is to say:

$$\Sigma_c' = h^2 * \begin{bmatrix} w_1^2\sigma_{c1}^2 & & \\ & \ddots & \\ & & w_d^2\sigma_{cd}^2 \end{bmatrix} \tag{14}$$

From the intuitive viewpoint, a good w should make q(x) for the transformed dataset D_y to approximate p(x) for the original dataset D_X as well as possible. Therefore, we require a criterion to evaluate w. ISE criterion (Girolami and Chao, 2003; Wang, 2008) has been investigated as an error criterion which will be less influenced by the presence of noise and outliers in the dataset. ISE criterion is a measure of the global accuracy of a density estimate, which converges to the mean squared error asymptotically. For the above two density estimates, w which provides the minimum of ISE is as follows:

$$\begin{aligned} w &= \arg\min_w \mathrm{ISE}(w) \\ &= \arg\min_w \int_{X^d} |p(x) - q(x)|^2\, dx \\ &= \arg\min_w \left(\int_{X^d} q^2(x)dx - 2\int_{X^d} p(x)q(x)dx\right) \end{aligned} \tag{15}$$

where, the term $\int_{X^d} p^2(x)dx$ has been dropped from the above since it is not dependent on the weight vector w. We should keep in mind that when we attempt to use ISE criterion for unsupervised feature ranking, MI criterion used in the Torkkola's work about supervised feature extraction (Torkkola, 2003) actually provides us alternative possible choice. However, in terms of the framework of the Torkkola's (2003) study, except for $\int_{X^d} q^2(x)dx$ and $\int_{X^d} p(x)q(x)dx)$, the term $\int_{X^d} p^2(x)dx$ still needs to be computed in MI criterion. Therefore, using ISE criterion here may result in further reducing the computational burden, compared with Torkkola (2003). In other words, the computational merit motivates us to use ISE criterion instead of MI criterion in this study.

Let us derive $\int_{X^d} q^2(x)dx$ and $\int_{X^d} p(x)q(x)dx)$ now (Dash *et al*., 1997; Torkkola, 2003):

$$\int_{X^d} G(x - x_i, \Sigma_1)G(x - x_j, \Sigma_2)dx = G(x_i - x_j, \Sigma_1 + \Sigma_2) \tag{16}$$

we have

$$\begin{aligned}\int_{X^d} q^2(x)dx &= \frac{1}{N^2}\int_{X^d}(\sum_{c=1}^{N_c}\sum_{i=1}^{J_c}\sum_{l=1}^{N_c}\sum_{j=1}^{J_l} G(x - y_{ci}, \Sigma_c^{'})G(x - y_{lj}, \Sigma_l^{'}))dx \\ &= \frac{1}{N^2}\sum_{c=1}^{N_c}\sum_{i=1}^{J_c}\sum_{l=1}^{N_c}\sum_{j=1}^{J_l}\int_{X^d} G(x - y_{ci}, \Sigma_c^{'})G(x - y_{lj}, \Sigma_l^{'}))dx \\ &= \frac{1}{N^2}\sum_{c=1}^{N_c}\sum_{i=1}^{J_c}\sum_{l=1}^{N_c}\sum_{j=1}^{J_l} G(y_{ci} - y_{lj}, \Sigma_c^{'} + \Sigma_l^{'}) \\ &= \frac{1}{N^2}\sum_{c=1}^{N_c}\sum_{i=1}^{J_c}\sum_{l=1}^{N_c}\sum_{j=1}^{J_l} G(w \otimes (x_{ci} - x_{lj}), \Sigma_c^{'} + \Sigma_l^{'})\end{aligned} \tag{17}$$

Similarly, we have:

$$\begin{aligned}\int_{X^d} p(x)q(x)dx &= \int_{X^d}\frac{1}{N^2}(\sum_{c=1}^{N_c}\sum_{i=1}^{J_c}\sum_{l=1}^{N_c}\sum_{j=1}^{J_l} G(x - y_{ci}, \Sigma_c^{'})G(x - x_{lj}, \Sigma_l))dx \\ &= \frac{1}{N^2}\sum_{c=1}^{N_c}\sum_{i=1}^{J_c}\sum_{l=1}^{N_c}\sum_{j=1}^{J_l}\int_{X^d} G(x - y_{ci}, \Sigma_c^{'})G(x - x_{lj}, \Sigma_l)dx \\ &= \frac{1}{N^2}\sum_{c=1}^{N_c}\sum_{i=1}^{J_c}\sum_{l=1}^{N_c}\sum_{j=1}^{J_l} G(y_{ci} - x_{lj}, \Sigma_c^{'} + \Sigma_l) \\ &= \frac{1}{N^2}\sum_{c=1}^{N_c}\sum_{i=1}^{J_c}\sum_{l=1}^{N_c}\sum_{j=1}^{J_l} G(w \otimes x_{ci} - x_{lj}, \Sigma_c^{'} + \Sigma_l)\end{aligned} \tag{18}$$

THE DETAILS OF THE PROPOSED APPROACH

Since, the weight vector w reflects the importance of every feature in the dataset, thus, the corresponding feature ranking can be achieved using the obtained w. Presentgoal is to find out w such that:

$$\begin{aligned} w &= \arg\min_{w} ISE(w) \\ &= \arg\min_{w} \frac{1}{N^2}\left\{\sum_{c=1}^{N_c}\sum_{i=1}^{J_c}\sum_{l=1}^{N_c}\sum_{j=1}^{J_l} G(w \otimes (x_{ci} - x_{lj}), \Sigma_c^{'} + \Sigma_l^{'}) \right. \\ &\quad \left. -2\sum_{c=1}^{N_c}\sum_{i=1}^{J_c}\sum_{l=1}^{N_c}\sum_{j=1}^{J_l} G(w \otimes x_{ci} - x_{lj}, \Sigma_c^{'} + \Sigma_l)\right\}\end{aligned} \tag{19}$$

Where:

$$w = (w_1, w_2, \cdots, w_d)^T, \sum_{i=1}^{d} w_i = d$$

The gradient descent method is adopted in the proposed approach. First, we need the derivatives of ISE(w), which are computed as follows.
Since,

$$\frac{\partial}{\partial w_k} G(w \otimes (x_{ci} - x_{lj}), \Sigma_c^{'} + \Sigma_l^{'}) = -\frac{G(w_k(x_{cik} - x_{ljk}), \Sigma_c^{'} + \Sigma_l^{'})}{w_k} \tag{20}$$

$$\begin{aligned}&\frac{\partial}{\partial w_k} G(w \otimes x_{ci} - x_{lj}, \Sigma_c^{'} + \Sigma_l) \\ &\quad = -\frac{(w_k x_{cik} - x_{lk})(\sigma_{ck}^2 w_k x_{ljk} + \sigma_{lk}^2 x_{cik})}{(w_k^2 \sigma_{ck}^2 + \sigma_{lk}^2)^2} G(w \otimes x_{ci} - x_{lj}, \Sigma_c^{'} + \Sigma_l)\end{aligned} \tag{21}$$

in terms of Eq. 19-21, we immediately have:

$$\frac{\partial ISE(w)}{\partial w_k} = \frac{-1}{N^2}\sum_{c=1}^{N_c}\sum_{i=1}^{J_c}\sum_{l=1}^{N_c}\sum_{j=1}^{J_l}\left[\begin{array}{l}\dfrac{G(w_k(x_{cik} - x_{ljk}), \Sigma_c^{'} + \Sigma_l^{'})}{w_k} - \\ \dfrac{2(w_k x_{cik} - x_{lk})(\sigma_{ck}^2 w_k x_{ljk} + \sigma_{lk}^2 x_{cik})}{(w_k^2 \sigma_{ck}^2 + \sigma_{lk}^2)^2} \\ G(w \otimes x_{ci} - x_{lj}), \Sigma_c^{'} + \Sigma_l) \\ \end{array}\right] \tag{22}$$

Thus, we further have the following learning rule for w:

$$w_k(t+1) = w_k(t) - \alpha\frac{\partial ISE(w)}{\partial w_k} \quad k = 1, 2, \cdots, d \tag{23}$$

where, α is the learning rate which is taken as 0.1 in our study.

Before giving the complete description of the proposed approach, we must discuss a particular implementation detail. As is seen before, when weighting all the features, we employ the constraint:

$$w = (w_1, w_2, \cdots, w_d)^T, \sum_{i=1}^{d} w_i = 1, i = 1, 2, \cdots, d$$

Such a constraint will pose a subtle issue, i.e., solving

$$\arg\min_{w} ISE(w)$$

with this constraint will yield all the same $w_i(i = 1,2,...,d)$, i.e., $w_1 = w_2 = ... =w_d = 1/d$. There are two strategies which can be used to circumvent this trouble. Suppose we have prior knowledge about the least important feature (e.g., the ith feature), the first strategy is to force $w_i = 0$ during the whole running period of the proposed approach. If we do not have such prior knowledge, the second strategy will start to work. We first take an arbitrary feature (e.g., the ith feature) and let $w_i = 0$ during the whole running period and then run the proposed approach. Accordingly, we choose the least important feature (e.g., the jth feature) and let $w_j = 0$ during the whole running period, then run the proposed approach again. In this way, we can easily obtain the final feature ranking for the dataset to be analysed. Please note, since we are interested in important features instead of the least important ones in feature ranking, such an arrangement based on the second strategy has almost no impact on the final feature ranking.

For the sake of the space of the paper, assume the first strategy is taken, let us state the proposed approach as follows:

```
Step 1: Given the dataset D, initialize $w = (w_1,w_2,...,w_d)^T$
        and let $w_i$ corresponding to the least important
        feature be zero; $t = 0$
Step 2: If all the data points in D have been classified
        with their class labels, skip this step. Otherwise,
        determine the number of classes and use the
        fuzzy clustering algorithm FCM (Dunn, 1973;
        Bezdek, 1981) to partition D such that the
        modified Parzen window density estimator in (10)
        can be employed
Step 3: Compute ISE using Eq. 19
Step 4: Compute $w_k$ (t+1) using Eq. 23
Step 5: $t \leftarrow t+1$
Step 6: If ISE reaches its minimum or t achieves the
        predefined maximum, then go to step 7, otherwise
        go to step 3
Step 7: Output the ranking result of features, according
        to the obtained w
```

EXPERIMENTAL RESULTS

We present the experimental results for eleven benchmarking datasets to validate the effectiveness of the proposed approach. It should be emphasized that we do not claim that the experimental results of present approach here are superior to current feature-ranking approaches. They show its power as a classifier-free approach by (1) comparing it with another classifier-free approach based on the traditional Parzen window density estimator and (2) experimentally proving it to be comparable to other supervised ranking approach.

Three groups of experiments were carried out with MATLAB 6.0 on the computer of Pentium IV with 2.66 GHz CPU and 512 MB memory. The first group of experiments deals with eleven datasets where algorithm FCM was first used to partition these datasets and then the proposed approach was executed to obtain the corresponding ranking results of features contained in these datasets. We verified the obtained ranking results using the k-fold cross-validation test (Richard, 2000) and the LIBSVM tool (Chang and Lin, 2001) and did a comparative study with other approaches. In k-fold cross-validation test, a dataset is divided into k subsets. Each time, one of the k subsets is used as the test set and all other k-1 subsets are put together to form a training set. Then the average accuracy or error across all trials is computed. In the second group of experiments, three datasets are involved, where the datasets Pipeline and Landsat contain their test data and the dataset WBC (Wisconsin Diagnostic Breast Cancer) does not have the test data. For the former, the classification accuracy, as the performance index, is used to evaluate the proposed approach. For the latter, the cross-validation test is adopted. In the third group of experiments, we presented a comparative result for the dataset Vowel with the latest advance (Andonie and Cataron, 2005).

Group A: This group of experiments deals with eleven UCI (Blake and Merz, 1998) datasets, as shown in Table 1. For the datasets Iris, New-Thyroid, Pima-Diabetes, Breast-Cancer, Wine, Ionosphere and Sonar, Table 2 shows the obtained feature ranks using the proposed approach and SUD approach (Dash *et al*., 1997), RELIEF approach (Kira and Rendell, 1992; Kononenko, 1994; Dash *et al*., 1997). Table 3 demonstrates the obtained feature ranks for the dataset Pima-Diabetes using the above three approaches and K-means-based approach (Girolami and Chao, 2003). The obtained feature ranks for the dataset wine is reported in Table 4 using both the above four approaches and FSM approach (Law *et al*., 2002). Table 5 demonstrates the feature ranks for the dataset Sonar using the proposed approach and RELIEF approach. Please note, in these tables, the relevant results of SUD and RELIEF are drawn (Dash *et al*., 1997), those of K-means-based approach are (Girolami and Chao, 2003) and those of FSM are from Law *et al*. (2002).

Here, let us briefly introduce the basic ideas of these approaches. SUD is based on the observation that removing an irrelevant feature from the feature set may not change the underlying concept of the data, but not so

Table 1: Eleven UCI datasets

Name	No. of dimensions	No. of data points		No. of classes
Iris	4	150		3
New-Thyroid	5	215		3
Pima-Diabetes	8	768		2
Breast-Cancer	9	699		2
Wine	13	178		3
Ionosphere	34	351		2
Sonar	60	208		2
WBC	30	569		2
Pipeline	12	1000 (training)	1000 (test)	3
Landsat	36	4435 (training)	2000 (test)	6
Vowel	10	528 (training)	462 (test)	11

Table 2: Feature ranks using three approaches

Datasets	The proposed approach	SUD	RELIEF-F
Iris	3,4,2,1	3,4,1,2	4,3,1,2
New-Thyroid	5,4,3,2,1	4,5,3,2,1	4,3,1,2,5
Breast-Cancer	2,8,6,4,3,5,1,7,9	1,7,3,2,5,6,4,8,9	6,2,3,7,5,1,4,8,9
Ionosphere	13,15,11,19,21,23,17,10,1,29,12,7,8,6,5,14 18,4,2,4,16,30,20,34,33,22,28,32,26,31,3,9 25,27,2	13,15,11,9,7,17,19,21,5,3,23,25,27,29, 31,33,10,4,6,12,14,8,16,20,18,22,28,26 24,30,32,34,2,1	34,22,33,6,4,8,16,14,21,9,27,15,30,20, 29,24,32,7,12,18,10,11,3,5,28,25,26,19 23,1,31,13,17,2

Table 3: Feature ranks using four approaches for Pima-Diabetes

The proposed approach	SUD	RELIEF-F	K-means-based approach
8,1,2,5,7,4,6,3	8,5,1,3,6,4,7,2	8,1,2,5,6,4,7,3	8,4,3,1,2,6,7,5

Table 4: Feature ranks using five approaches for Wine

The proposed approach	SUD	RELIEF-F	FSM	K-means-based approach
13,7,12,2,11,8,3,1,6,10,4,5,9	7,6,12,9,11,10,5,13,1,4,3,8,2	6,9,1,11,5,7,10,4,12,2,13,3,8	7,12,6,1,9,11,10,13,2,8,4,5,3	6,7,12,9,11,10,5,13,1,4,3,8,2

Table 5: Feature ranks using the proposed approach and RELIEF-F for Sonar

The proposed approach	45,1,11,52,51,12,10,54,48,4,2,13,49,46,36,9,21,58,3,20,5,35,53,19,44,34,47,22,18,29,17,31,30,28,37,33,16,56,32,60,15,14,43,59, 23,56,27,6,24,26,50,25,8,42,57,7,38,39,40,41
RELIEF-F	45,12,11,49,44,10,9,48,46,13,54,47,1,36,21,43,2,52,20,35,28,37,51,4,55,8,59,22,5,6,27,16,58,50,14,52,31,34,19,3,17, 53,32,23,7,40,39,15,60,29,33,24,56,25,30,26,41,18,38,57

otherwise. Each time a relatively unimportant feature is removed from the feature set by using the entropy index (Dash *et al*., 1997). The adopted RELIEF-F is a modification of the original RELIEF (Kononenko, 1994; Dash *et al*., 1997), a key idea of which is to estimate the quality of features according to how well their values distinguish between data points that are near to each other. The original RELIEF can deal with nominal and numerical features. However, it cannot deal with incomplete data and is limited to two-class problems. Its extended version RELIEF-F solves these problems. The feature ranking approach based on K-means clustering involves classification capabilities of feature vectors and correlation analysis between two features. It can rank the features by using the correlation index between two features (Girolami and Chao, 2003). FSM approach is based on a feature saliency measure which is obtained by an EM algorithm. However, it assumes that the features are conditionally independent, given the components. It ranks the features in descending order of saliency (Law *et al*., 2002).

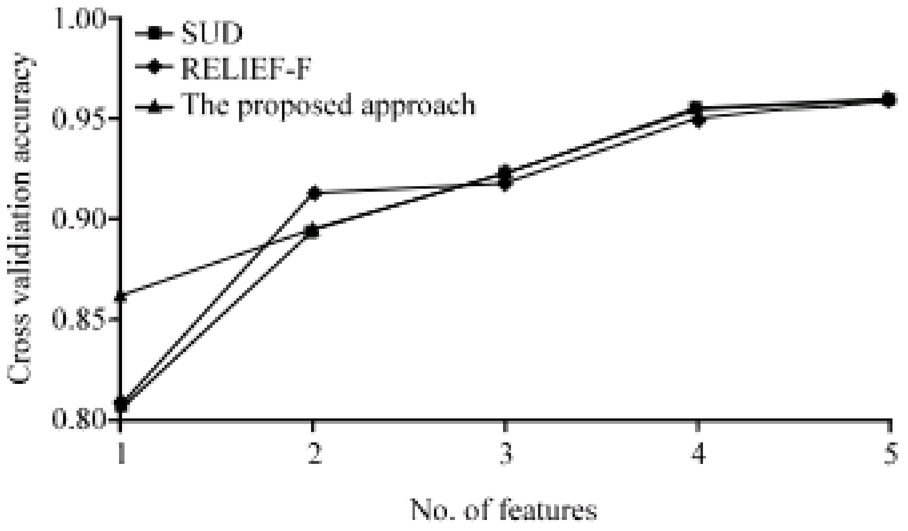

Fig. 1: Cross-validation for the dataset New-Thyroid

As we may know well, the most important features in Iris dataset are of the third and fourth. Just like SUD, RELIEF-F, the proposed approach yielded the same ranking result. Figure 1-6 demonstrate the 5-fold cross-validation accuracies for other datasets New-Thyroid, Pima-Diabetes, Breast-Cancer, Wine, Ionosphere and Sonar, which clearly indicate that in most cases, the proposed approach is comparativelycompetitive to other

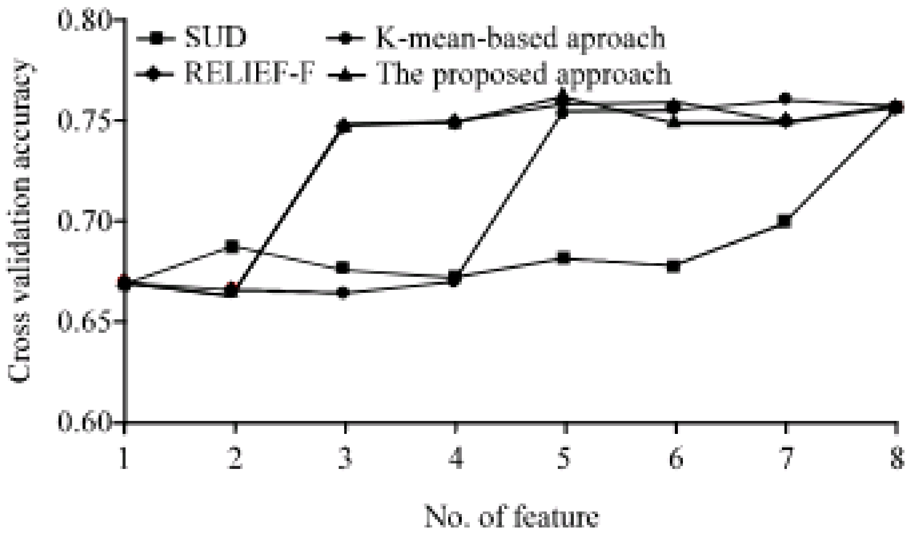

Fig. 2: Cross-validation for the dataset Pima-Diabetes

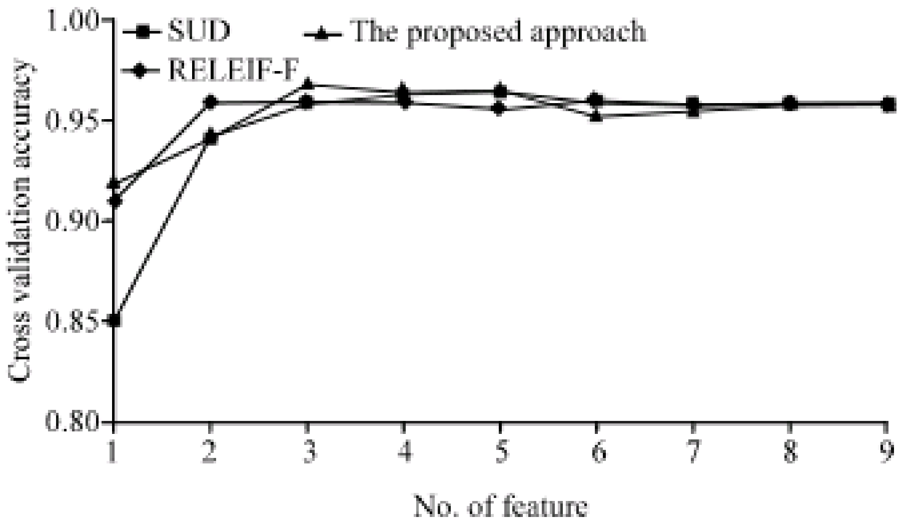

Fig. 3: Cross-validation for the dataset Breast-Cancer

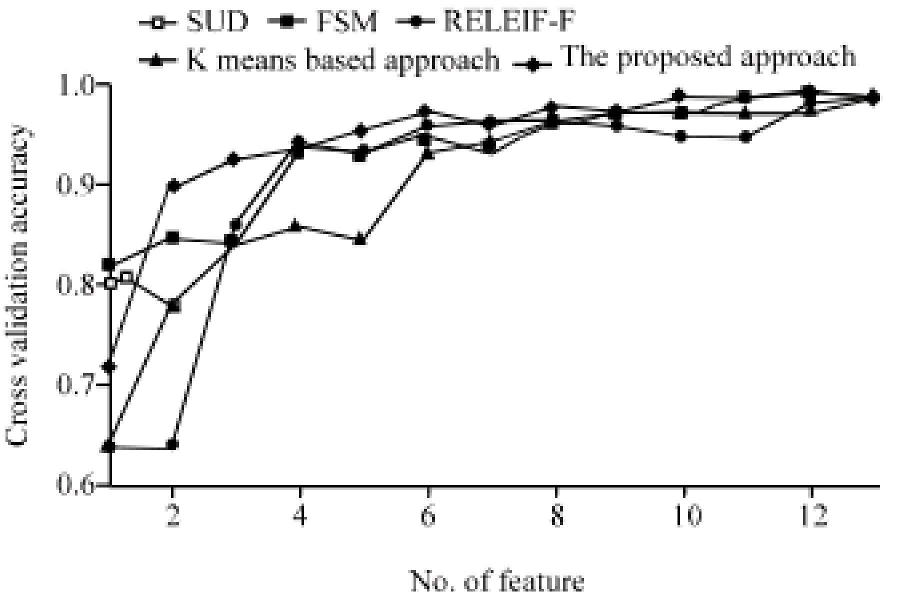

Fig. 4: Cross-validation for the dataset Wine

approaches. Figure 7a-d shows the visualization results for two most important features of the dataset Wine using these approaches. It is clear that the proposed approach had better visualization capability than other approaches, (Fig. 7d).

Of course, since the combinational computation of Gaussian functions is involved in the proposed approach, its computational time is still comparatively high. For example, it requires 4.438 sec for Iris dataset and 4.406 sec for New-Thyroid dataset, however, 39.453 for Pima-Diabetes dataset, 38.922 sec for Breast-Cancer dataset and 34.141 sec for Ionosphere dataset. This actually poses a noting-worthy issue, i.e., we should manage to further reduce its computational burden in near future.

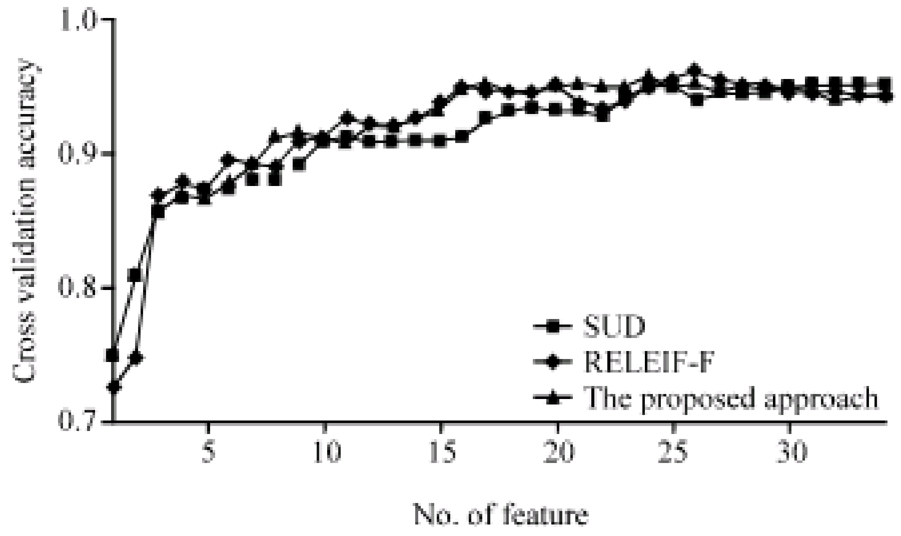

Fig. 5: Cross-validation for the dataset Ionosphere

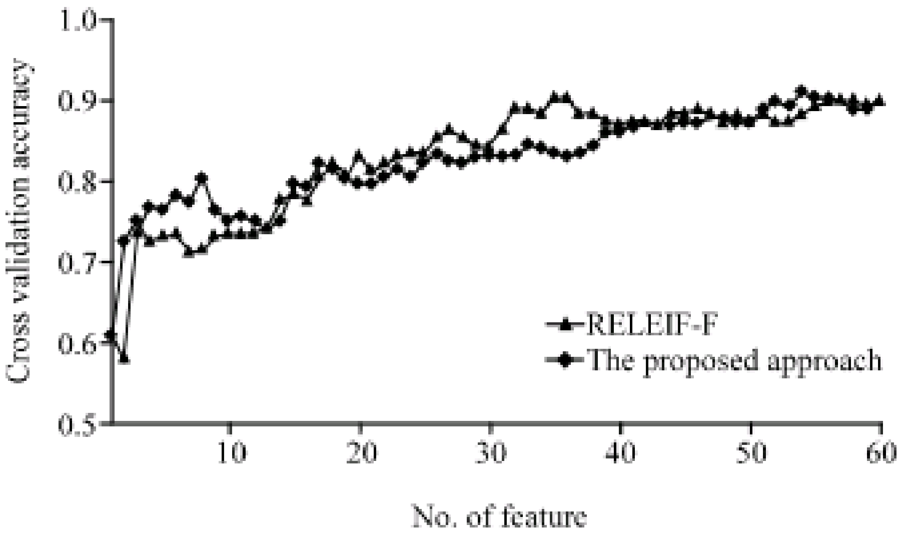

Fig. 6: Cross-validation for the dataset Sonar

Group B: This group of experiments deals with three datasets: Pipeline, Landsat and WBC, as shown in Table 1. All the data in these three datasets have been classified using the class labels. Table 6 demonstrates the feature ranks for these three datasets using the proposed approach. For the datasets Pipeline and Landsat, we employed LIBSVM (Chang and Lin, 2001) to their training data according to the obtained feature ranks and then obtained the classification accuracies for their test data, (Fig. 8-9). Figure 10 demonstrates the 5-fold cross-validation accuracies for the dataset WBC. In Fig. 10 and Table 6, a comparison for this dataset between the modified Parzen window density estimator and the traditional one is also included. In particular, two very different feature ranking are generated by traditional Parzen window density estimator and the modified Parzen window density estimator, Table 6. This just show the inappropriateness of the traditional Parzen window density estimator. We can see the effectiveness of the proposed approach from these Fig. 10 and Table 6. That is to say, with the proposed approach, only a small number of the ranked features are taken, the corresponding comparatively high classification/cross-

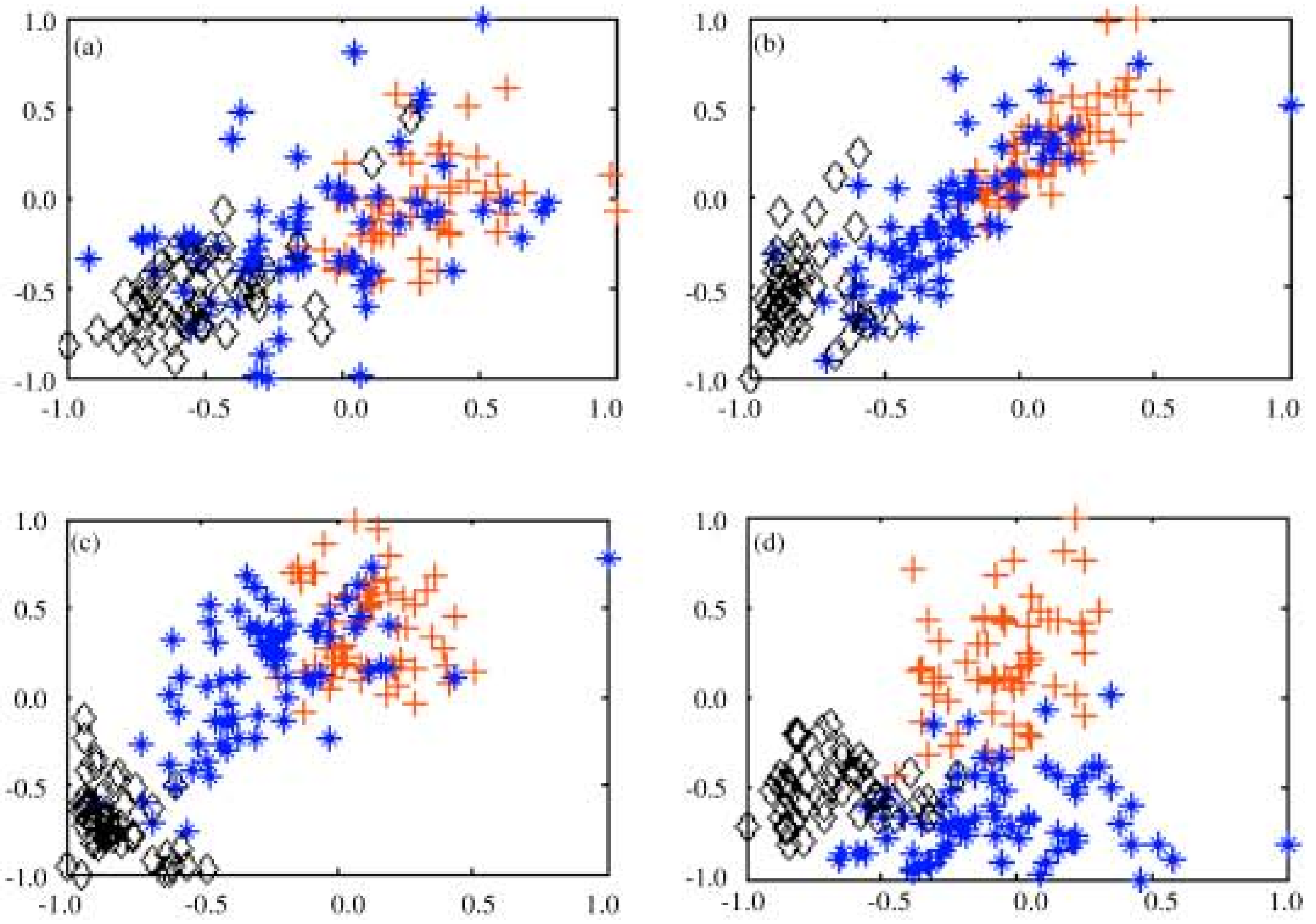

Fig. 7: Visualization results using the five approaches for the two most important features of the dataset Wine, (a) RELIEF-F approach, (b) SUD approach and K-means-based approach, (c) FSM approach and (d) The proposed approach

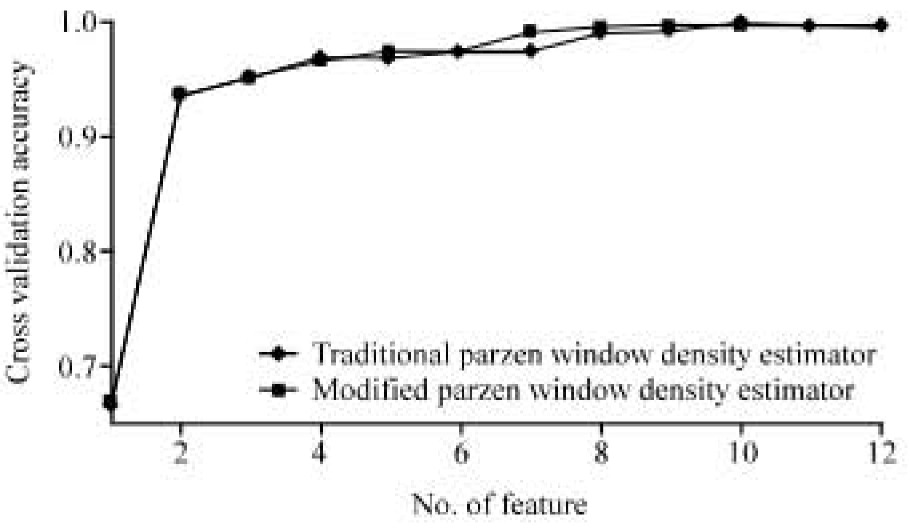

Fig. 8: Classification for the dataset Pipeline using LIBSVM

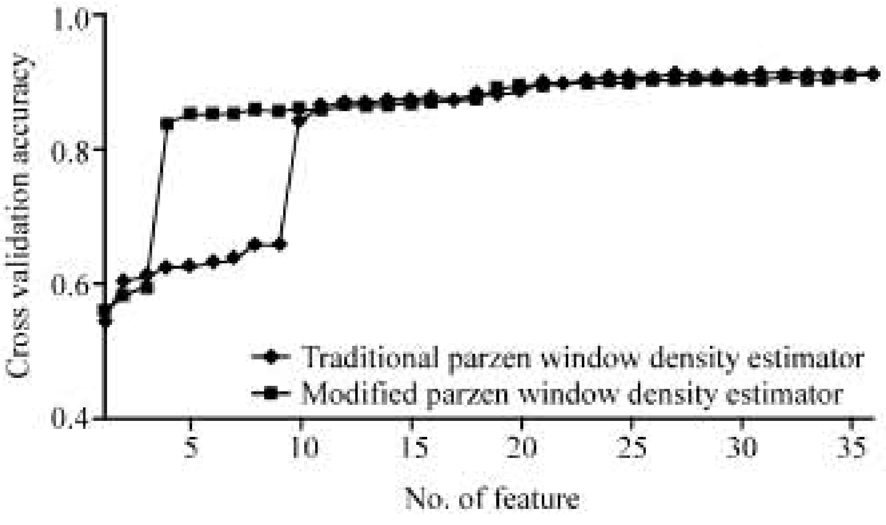

Fig. 9: Classification for the dataset Landsat using LIBSVM

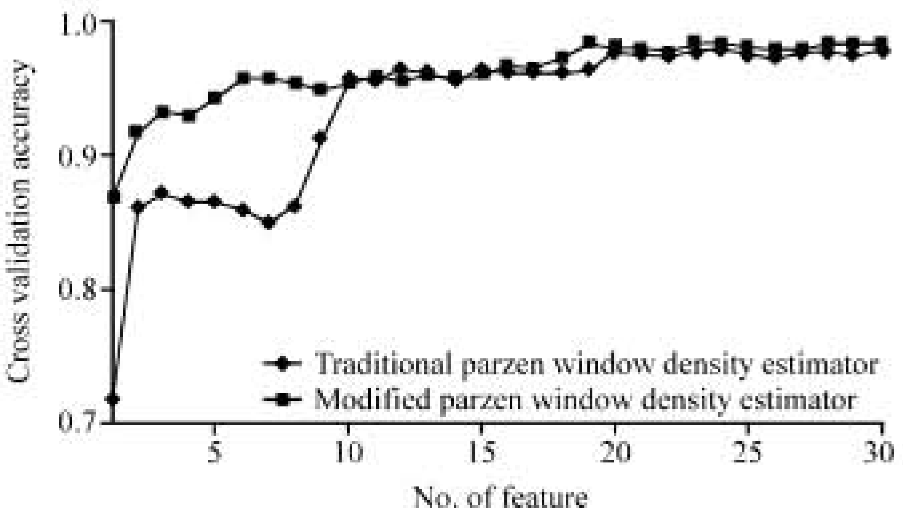

Fig. 10: Cross-validation for the dataset WBC using LIBSVM

validation accuracies for these three datasets were obtained. For example, we presented 2-5 ranked features from the dataset Pipeline obtained by the proposed approach, to LIBSVM, the obtained classification accuracies for the test data are 93.4, 94.9, 96.3 and 97.1.

Group C: Only the data set Vowel is involved here to exhibit the comparative study between the proposed approach and other feature ranking approaches RLVQ, GRLVQ, SRNG, ERLVQ and ESRNG in the latest literature (Andonie and Cataron, 2005). As we may know well, Standard Learning Vector Quantization (LVQ) does not

Table 6: Feature ranks using the proposed approach for Pipeline, Landsat, WBC

Pipeline	Traditional Parzen window density estimator	7,1,11,10,9,5,2,8,12,6,3,4
	Modified Parzen window density estimator	11,1,7,5,3,6,9,4,10,2,12,8
Landsat	Traditional Parzen window density estimator	17,13,21,18,29,5,1,33,9,25,22,30,14,34,6,2,10,26,19,23,27,24,20,28,15,11,35,31,12,16,32, 7,3,36,8,4
	Modified Parzen window density estimator	33,1,13,25,29,21,17,9,5,11,27,3,7,35,23,15,31,12,19,28,10,4,36,34,16,8,26,24, 32,22,2,20,30,14,6,18
WBC	Traditional Parzen window density estimator	14,24,8,13,4,23,11,21,28,3,7,1,26,6,29,22,27,2,25,9,5,30,12,19,16,15,18,10,20,17
	Modified Parzen window density estimator	17,4,20,13,11,15,19,12,30,24,16,29,28,4,10,26,27,7,6,2,8,9,23,21,3,1,28,22,5,25

Table 7: Feature ranks using seven approaches for Vowel

RLVQ	GRLVQ	SRNG	ERLVQ	EGRLVQ	ESRNG	The proposed approach
2,5,1,9,6,3,4,8,7,10	2,5,4,6,3,1,9,7,8,10	1,4,6,2,3,9,8,5,7,10	2,1,3,4,6,8,9,5,10,7	3,1,2,6,5,4,9,8,7,10	2,1,3,8,9,4,5,10,8,7	1,2,8,5,3,6,4,10,9,7

Table 8: Classification accuracies using LIBSVM for test data of Vowel (%)

	No. of features									
Algorithm	1	2	3	4	5	6	7	8	9	10
RLVQ	36.3	36.1	50.4	56.7	52.6	51.9	60.8	61.0	58.0	59.9
GRLVQ	36.3	36.1	33.3	40.6	47.8	60.1	60.8	55.4	58.0	59.9
SRNG	27.0	29.2	35.2	48.7	56.7	59.3	58.6	61.0	58.0	59.9
ERLVQ	36.3	51.3	53.7	59.0	56.7	60.3	58.6	61.0	62.5	59.9
EGRLVQ	9.3	25.1	53.7	52.6	53.4	60.1	60.8	61.0	58.0	59.9
ESRNG	36.3	51.3	53.7	54.3	54.3	60.3	61.4	58.6	62.5	59.9
The proposed approach	27.0	51.3	50.4	55.1	54.9	51.3	61.2	62.1	62.5	59.9

discriminate between more or less informative features: their influence on the distance function is equal. On the contrary, Relevance LVQ (RLVQ) in (Bojer *et al*., 2001; Andonie and Cataron, 2005) holds a changeable relevance value for every feature and employs a weighted distance function for classification. An iterative heuristic training process is used to tune the weight values for a specific problem: the influence of features which frequently contribute to misclassifications of the system is reduced while the influence of very reliable features is increased. Generalized RLVQ (GRLVQ) is the modification of RLVQ. This approach modifies RLVQ by using an adaptive metric and leads to a more powerful classifier with little extra cost compared with RLVQ. The Supervised Relevance Neural Gas (SRNG) approach combines the neural-gas (NG) algorithm (Hammer *et al*., 2005) and GRLVQ. The idea was to incorporate neighborhood cooperation of NG into GRLVQ to speedup the convergence and make initialization less crucial. Energy SRNG (ESRNG) approach uses the maximization of the informational energy (IE) as a criterion for computing the relevancies of input features. This adaptive relevance determination is used in combination with the SNG model. Energy Generalized Relevance LVQ (EGRLVQ) approach uses the estimation of the Informational Energy (IE) as a maximization criterion for computing the feature relevance. Both ESRNG and EGRLVQ measure the informational energy using Onicescu's informational energy (Cataron and Andonie, 2004).

Table 7 demonstrates the experimental results using both the above approaches and the proposed approach here. We directly draw the results of all other methods except the proposed approach from (Andonie and Cataron, 2005) into this Table 7. Based on the obtained feature ranks, LIBSVM is first applied to the training data of this dataset and then to its test data. As seen in the Table 8, when 2-5 ranked features are taken, the obtained classification performance using the proposed approach is comparable to ESRNG and a little less than ERLVQ. For other cases except one or six feature cases, the proposed approach is generally comparable to other approaches. However, please let us keep in mind the fact that the proposed approach attains these comparable results in the case where it is classifier-free while other approaches here are supervised. What is more, since all approaches yield very low classification accuracies (much less than 50%) for one feature case, it is unnecessary for us to analyze this case. For the six feature case in Table 8, although the proposed approach obtains the lowest classification accuracy, it is still a little bigger than 50% (i.e., 51.3%). In summary, the proposed approach is effective in most cases for the dataset Vowel.

CONCLUSIONS

This study describes a novel feature ranking approach which may be attractive for pattern recognition tasks on high-dimensional datasets. The proposed approach provides three major contributions. First, the proposed Parzen window density estimator takes non-uniform distribution contained in the datasets into account, so the proposed estimator can make use of more information of the datasets than the conventional Parzen window density estimator. Second, instead of the commonly used MI based criterion, the novel ISE criterion, as an alternative way, is adopted in our study. Present experimental results demonstrate that the proposed approach based on ISE criterion has as least the same or comparable performance as current feature ranking approaches. Third, the proposed approach requires no prior knowledge of the datasets to be analyzed and it can be easily realized by using the corresponding gradient descent method.

Although the proposed approach is based on ISE criterion rather than MI criterion, it still has comparatively high computational burden due to the existence of the combinational computation of Gaussian functions. We will study how to further reduce its computational time, possibly using the random sampling method, in near future. Further research may include theoretical research on the proposed approach and explore more practical applications.

ACKNOWLEDGMENTS

This research is supported by 2007 National 863 project (Grant No. 2007AA1Z158), 2007 two grants (No. 60773206, 60704047) from National Science Foundation of China, New_century Outstanding Young Scholar Grant of Ministry of Education of China (NCET-04-0496), National KeySoft Lab. at Nanjing University, The Key Lab. of Computer Information Technologies at JiangSu Province, the 2004 key project of Ministry of Education of China and National Key Lab. of Pattern Recognition at Institute of Automation, CAS, China.

REFERENCES

Andonie, R. and A. Cataron, 2005. Feature ranking using supervised neural gas and informational energy. Proceedings of International Joint Conference on Neural Networks, July 31-Aug. 4, Montreal, Canada, pp: 1269-1273.

Basak, J., R.K. De and S.K. Pal, 1998. Unsupervised feature selection usinga neuro fuzzy approach. Pattern Recognition Lett., 19: 997-1006.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algoritms. 1st Edn., Plenum Press, New York, ISBN: 0306406713.

Biesiada, J., W. Duch, A. Kachel, K. Maczka and S. Palucha, 2005. Feature ranking methods based on information entropy with parzen windows. Proceedings of the 9th International Conference on Res. in Electrotechnology and Applied Informatics (REI), Aug. 31-Sep. 3, Katowice-Kraków, Poland, pp: 109-119.

Bishop, C., 1995. Neural Networks for Pattern Recognition. 1st Edn. Oxford University Press, USA, ISBN-13: 978-0198538646.

Blake, C.L. and C.J. Merz, 1998. UCI Repository of Machine Learning Databases. 1st Edn., University of California, Irvine, CA.

Bojer, T., B. Hammer, D. Schunk and K.T.V. Toschanowitz, 2001. Relevance determination in learning vector quantization. Proceedings of the European Symposium on Artificial Neural Networks (ESANN), Apr. 2001, D-Side Publications, pp: 271-276.

Cataron, A. and R. Andonie, 2004. Energy generalized LVQ with relevance factors. Proceedings of the IEEE International Joint Conference on Neural Networks IJCNN, July 2004, IEEE Computer Society, pp: 1421-1426.

Chang, C.C. and C.J. Lin, 2001. LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Dash, M., H. Liu and J. Yao, 1997. Dimensionality Reduction of Unsupervised Data. Newport Beach. Proceedings of 9th IEEE International Conference Tools with Artificial Intelligence, Nov. 1997, IEEE Computer Society, pp: 532-539.

Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cyber., 3: 32-57.

Dy, J.G. and C.E. Bradley, 2004. Feature selection for unsupervised learning. J. Mach. Learning Res., 5: 845-889.

Girolami, M. and H. Chao, 2003. Probability density estimation from optimally condensed data samples. Probability density estimation from optimally condensed data samples. IEEE Trans. Pattern Anal. Mach. Intell., 25: 1253-1264.

Guyon, I. and A. Elisseeff, 2003. An introduction to variable and feature selection. J. Mach. Learn. Res., 3: 1157-1182.

Hammer, B., C. Information, M. Strickert and T. Villmann, 2005. Supervised neural gas with general similarity measure. Neural Process. Lett., 21: 21-44.

Huang, D. and T. Chow, 2003. Searching optimal feature subset using mutual information. Proceedings of the Eur. Symposium on Artificial Neural Networks (ESANN), Aug. 2003, d-side publications, pp: 161-166.

Kira, K. and L. Rendell, 1992. A practical approach to feature selection. Proceedings of the 9th Int. Workshop on Machine Learning, July 1992, California: Morgan Kaufmann, pp: 249-256.

Kononenko, I., 1994. Estimating attributes: Analysis and extension of RELIEF. Proceedings of Eur. Conference on Machine Learning, Aug. 1994, Springer Berlin, pp: 171-182.

Law, M.H., A. Jain and M. Figueiredo, 2002. Feature selection in mixture-based clustering. Proceedings of Advances in Neural Information Processing Systems, Dec. 2002, The MIT Press, pp: 609-616.

Morita, M., R. Sabourin, F. Bortolozzi and C.Y. Suen, 2003. Unsupervised feature selection using multi objective genetic algorithms for handwritten word recognition. International Conference on Document Anal. and Recognition (ICDAR), August, IEEE Computer Society pp: 666-671.

Principe, J., D. Xu and J. Fisher, 2000. Information Theoretic Learning. In: Unsupervised Adaptive Filtering, Haykin, S. (Ed.). Wiley, New York, ISBN: 0471294128.

Richard, O.D., 2000. Pattern Classification. 2nd Edn., Wiley Interscience, New York, ISBN: 978-0-471-05669-0.

Song, F.X., X.M. Gao, S.H. Liu and J.Y. Yang, 2005. Dimensionality reduction with less loss in statistical pattern recognition. Chinese J. Computer, No.11 (In Chinese).

Torkkola, K., 2003. Feature extraction by non-parametric mutual information maximization. J. Mach. Learning Res., 3: 1415-1438.

Wang, H., D. Bell and F. Murtagh, 1999. Axiomatic approach to feature subset selection based on relevance. IEEE Trans. Pattern Anal. Mach. Intell., 21: 271-277.

Wang, S.T., F.L. Chung and F.S. Xiong, 2008. A novel image thresholding method based on Parzen window estimate. Pattern Recognit., 41: 117-129.