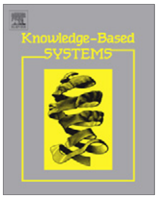
Knowledge-Based Systems 37 (2013) 94–104

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



# Feature selection based on cluster and variability analyses for ordinal multi-class classification problems

Hung-Yi Lin *

*National Taichung University of Science and Technology, No. 129, Sec. 3, San-min Road, Taichung, Taiwan*

ARTICLE INFO

*Article history:*
Received 22 November 2011
Received in revised form 21 July 2012
Accepted 23 July 2012
Available online 11 August 2012

*Keywords:*
Feature selection
Huang-index
Cluster validity
Multivariate analysis
Multi-class classification

ABSTRACT

Feature selection is an essential problem for pattern classification systems. This paper studies how to provide systems with the most characterizing features for ordinal multi-class classification task. The integration of cluster analyses and variability analyses advances a novel feature selection scheme with efficiency. The Huang-index method using fuzzy *c*-means is employed to enhance cluster validity and optimizes a consistent number of clusters among the features. A new entropy-based feature evaluation method is formulated for the authentication of relevant features. Then, multivariate statistical analyses are utilized to solve the redundancy between relevant features. Experimental results show that our new feature selection scheme sifts successfully a compact subset of characterizing features for classification problems with multiple classes.

## 1. Introduction

Classification problems have become more complex and intricate in modern digital applications when facing incessant data explosion [11]. Present databases are capable of enduring various data types. The multi-dimensionality of data entries makes a single record accommodate a great number of fields (attributes, features). In addition to these challenges, modern classification problems frequently hesitate to make decisions among a variety of classes. The ever-increasing growth in data complexity and class multiplicity has severely deteriorated classification performance and accuracy rate. To mitigate this situation, selective sampling and feature reduction are two potent strategies for simplifying classification tasks. The goal of selective sampling is to prevent outlier or noisy data entries from being involved in analytical and computational processes [26]. However, in many applications, the discrimination between typical and non-typical instances is ambiguous. Other problems can come from the insufficient amount of instances for sampling or the size of sampling data remains tremendous. Hence, feature reduction becomes an alternative, especially when the instance dimensionality is as big as hundreds or, even more, as thousands. Combination of several homogenous data sources or collaborating different datasets derived from the same theme can constitute high-dimensional datasets with ease. Feature reduction takes the responsibility in generating a smaller subset of essential classification factors. Although many well-designed feature reduction schemes have been proposed, no study has been made on how to simultaneously optimize necessary and sufficient factors in classification problems.

To authenticate precisely discriminative features from others, proper data preprocessing is important. Good data preprocessing can refine the discriminative power of features while improper data preprocessing are likely to cause failure in discovering pivotal features. Generally, variety makes continuous data more controversial than discrete (nominal or categorical) data. Variety simplification for continuous data is commonly required before executing feature evaluation. Discretization is the process of transferring continuous data into discrete counterparts. Discretization can preserve the original data distribution. However, for classification problems, data discrimination is more attractive than their distribution. To enhance the discriminative power of features, better data preprocessing using cluster analysis [20,21] is first investigated in this paper.

The multiplicity of target classification variable is another issue worthy of discussion. Data with multiple categories burden the classifier and in turn bring more complex patterns to the classification task. Classifiers are thus requested to maintain more powerful discriminative capability when dealing with multi-class problems. The construction of classification models for a large number of competing classes [1,4,33,43] has drawn much attention. These tasks are not handled well by general-purpose learning methods and are usually addressed in an ad hoc fashion [7,25,53].

Our work in this paper focuses on three issues that have not been explored in earlier work. First, although cluster analyses have been widely used for many applications, no cluster analyses with

* Tel.: +886 4 2219 6769; fax: +886 4 2219 6161.
*E-mail address:* linhy@nutc.edu.tw

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.07.018

validation study have been employed to comprehend data composition for multi-class classification problems. Thus the first goal in this paper is to present an appropriate data clustering scheme for accurately preprocessing the raw feature values.

Second, we investigate data variation inside subsets and include such information into our new feature evaluation criterion. By means of such design, we can identify the most discriminative features for the task of multi-class classification.

Third, a novel variability analysis relying on statistical multivariate analysis is proposed and a new heuristic algorithm facilitates the selection of a compact subset of minimal-redundancy and maximal-relevance features. The final goal of this paper is to select the necessary and sufficient factors for the ordinal multi-class classification problems.

This paper is organized as follows. Section 2 reviews the related work including feature selection, cluster analyses, and cluster validity methods. The justification for the use of fuzzy *c*-means algorithm and Huang-index is also explained in this section. Section 3 presents the new feature evaluation criterion. Section 4 describes the proposed feature selection scheme with the design of variability analysis using statistical multivariate analysis. The reason why principle component analysis can solve redundancy problems is also explained in this section. Both experimental and analytical results are presented in Section 5. Finally, concluding remarks are given in the last section.

## 2. Related work

Feature reduction (or data dimensionality reduction, DDR) [5,10,47] is broadly categorized into feature transform (or feature extraction) and feature selection. Feature extraction can transform input data into a reduced representation set of factors. It is expected that the factor set will extract the relevant information from the input data to facilitate the classification task using this reduced representation instead of the full-size input. Feature selection [13,22,28,44,46], also known as variable selection, is the technique of removing nearly all irrelevant and redundant features from the original feature set. The selected features in the reduced set take the major responsibility of the classification work. Feature selection has many advantages such as alleviating the effect of the curse of dimensionality and enhancing generalization capability. In addition, feature selection can benefit classification performance by speeding up the learning process and improving model interpretability.

In feature selection, it is well known that the combinations of individual discriminative features do not necessarily lead to good classification performance [35,40,53]. It is difficult to collect a subset of relevant features with respect to the class concept when they are totally independent. Redundancy among features is very common. Existing feature selection methods mainly exploit two strategies to reduce such problem: *individual evaluation* and *subset evaluation*. *Individual evaluation* ranks features according to their importance in differentiating instances of different classes. It can only remove irrelevant features as redundant ones once features are ranked as similar. The conventional methods include information entropy [24,30] and the Gini index [31]. *Subset evaluation* searches for a compact subset of features that satisfies some goodness measure and can exclude irrelevant features as well as redundant ones. Mutual information [9,18,36] adopts this strategy. Executing individual evaluation usually involves a complete search and thus requires the computational complexity of $O(N)$ to generate candidate feature subsets, where $N$ is the cardinal of feature space. However, even greedy sequential search algorithms integrated with heuristic or random search strategy need at least $O(N^2)$ and at most $O(2^N)$ execution complexities for finding an optimal subset. Therefore, individual evaluation methods are more time-consuming than subset evaluation methods. For the sake of efficiency, a novel individual evaluation method integrated with a new heuristic feature selection scheme is presented in this paper.

Feature selection algorithms fall into three broad categories [38,49]: *filter*, *wrapper*, and *embedded* models. *Filter techniques* [8] assess feature relevance scores and features with low relevance scores are removed. The advantages of filter techniques are threefold. First, they easily scale to high-dimensional datasets. Second, they are computationally simple and fast. Third, they are independent of the classification algorithm. However, their common disadvantage is that they ignore feature dependencies, resulting in poor classification performance. *Wrapper methods* [19] search the space of possible feature subsets for obtaining the best evaluation of a specific subset of features. They are capable of embedding the model hypothesis within the feature subset search and tailored to a specific classification algorithm. However, as the space of feature subsets grows exponentially with the number of features, high computational cost is easily imposed on the classifier construction. Heuristic search methods are frequently used toward an optimal subset. Another drawback is that they have a higher risk of overfitting than filter techniques. *Embedded techniques* [14,48] search the combined space of feature subsets and hypotheses for the optimal subsets of features. Such methods have the advantage of including the interaction with the classification model. Although they are dependent of the classification algorithm, far lower computational cost is required than wrapper methods. Our scheme possesses the superiorities of model independence in filter technique and heuristic search in wrapper method.

The four typical data types used in feature values are categorical, ordinal, integer, and real. Real-valued features are usually richer in information than categorical- and integer-valued ones. For the sake of reducing the mess information and enhancing the discriminative power, real-valued features necessitate preprocessing prior to all classification procedures. Common preprocessing strategy adopts normalization methods such as $z$ score, $z*$ score, strictly standardized mean difference (SSMD) [54,55], SSMD$*$, and $t$ statistic. Knowing how to use these methods correctly is critical because misusing them can readily produce misleading results. However, in the scenario of unsupervised learning, normalization methods cannot elaborate their superiorities since the population parameters cannot be collected in advance. Therefore, such situation motivates us to study the effect of cluster analyses on real-valued features. As well known, cluster analyses can achieve the explorative task of assigning a set of data into clusters so that the data in the same cluster are more similar to each other than to those in other clusters. In *hard clustering*, data are divided into *distinct* clusters, where each data element belongs to exactly one cluster. In *fuzzy clustering* (also termed *soft clustering*), data elements can belong to more than one cluster. Fuzzy clustering provides higher flexibility and thus better analytical quality than hard clustering. Hence, we adopt the most widely used fuzzy clustering algorithm, i.e., fuzzy *c*-means (FCM) algorithm in this paper. The main advantages of FCM are its simplicity and speed which allows it to run on large datasets.

In order to identify an appropriate number of clusters, it is commonly necessary to first define a measure of similarity or proximity that will establish a rule for assigning patterns to the domain of a particular cluster centers (prototype). As a consequence, different paradigms such as partitional [12], hierarchical [53], spectral [39], density-based and mixture-modeling [29], have been proposed according to different communities. Cluster analyses with validation study on real-valued features can help to comprehend the composition of data entries. Two typical clustering questions are frequently addressed: (i) how many clusters are actually present in the data and (ii) how real or good is the clustering itself. The problem in finding an optimal number of clusters is called the cluster validity problem [2]. A number of clustering methods [21,32,34] and validation indices [23,50,51,56] have been proposed and successfully employed to solve this problem. The designs of FCM together with the cluster validity indexes (PBMF-index and Huang-index) are briefly explained in the following.

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a dataset, FCM [3] minimizes the following objective function:

$$J_m = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^m \|x_i - c_j\|^2, \quad 1 < m < \infty, \tag{1}$$

where $c$ is the number of clusters and $m$ is any real number greater than 1, $u_{ij}$ is the level of membership of $x_i$ in the cluster $j$, $x_i$ is the $i$th $d$-dimensional measured data, $c_j$ is the $d$-dimensional center of the cluster, and $\|*\|$ is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the membership $u_{ij}$ and the cluster centers $c_j$ updated by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}, \tag{2}$$

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^m \times x_i}{\sum_{i=1}^{N} u_{ij}^m}. \tag{3}$$

This iteration will stop when $\max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon$, where $\varepsilon$ is a termination criterion between 0 and 1, whereas $k$ is the number of iteration steps. This procedure converges to a local minimum or a saddle point of $J_m$.

PBMF-index is defined by $V_{PBMF} = \left(\frac{1}{c} \times \frac{E_1}{J_m} \times D_c\right)^2$, where $E_1 = \sum_{j=1}^{n}\|x_j - v\|$ with $v$ being the centroid of the dataset and $D_c = \max_{i,j=1}^{c}\|c_i - c_j\|$. The Huang-index method [20] is designed according to the PBMF-index function and rough set theory. This method not only yields a superior clustering capability than traditional PBMF-index method, but also achieves a reliable classification. The preprocessing that involves application of FCM to real-valued features and validation by the Huang-index is outlined in Fig. 1.

## 3. Feature evaluation

The designs of Gini impurity, information gain, and information gain ratio pay common attention to *data distribution*. In other words, when a feature is tested, the probability distribution of different classes is calculated and employed to measure the relevance with respect to the class concept. Gini impurity is computed by summing the probability of each item being chosen times the probability of a mistake in categorizing that item, while information gain and information gain ratio are calculated using Shannon's entropy. Shannon's entropy is a board and general concept which studies the amount of information in a transmitted message. The probability that a particular message has actually been transmitted is a measure of the amount of information contained in the message. Data distributions in multi-class problems are looser and in turn have more diverse results than those in two-class problems. Therefore, features evaluated according to the conventional entropy criterion can yield similar measurements, making it hard to distinguish explicitly their discriminative powers from the similar evaluation results. To remedy such situation, we simply provide the information of data variation inside every subset to the conventional entropy criterion. The purpose of this design is to simultaneously take the dissimilarity around subsets (via data distribution) and the aggregation inside subsets (via data variation) into account. The measure of data variation is simply conducted by statistical variance. A feature with good discrimination should be capable of aggregating those data belonging to the same group. Let $T$ be a training dataset, data variation measured from every subset $T_i$ using data variance $\sigma^2(T_i)$ is employed to enhance the traditional Shannon's entropy. Intuitively, a subset $T_i$ comprises all consistent data of both zero entropy $H(T_i)$ and variance $\sigma^2(T_i)$. We formulate the enhanced entropy (denoted as $EH$) as a new feature evaluation criterion as follows:

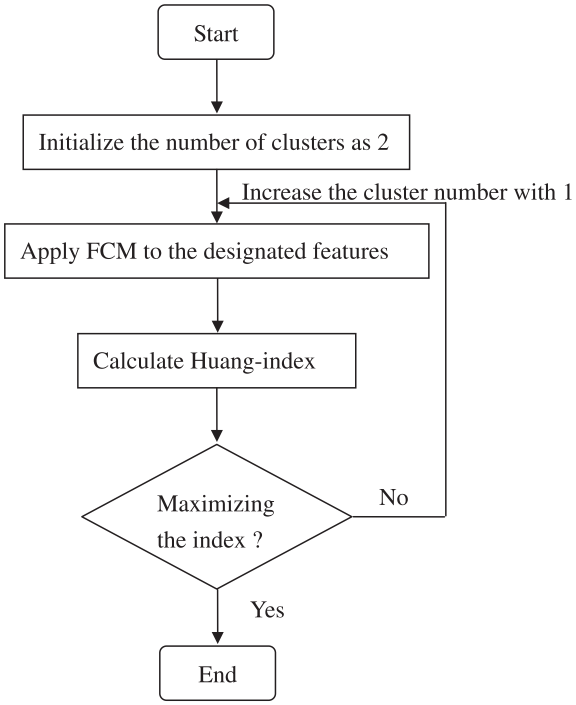

Start
Initialize the number of clusters as 2
Increase the cluster number with 1
Apply FCM to the designated features
Calculate Huang-index
Maximizing the index ?
No
Yes
End

**Fig. 1.** Flow chart of preprocessing.

$$EH(x, T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times H(x, T_i) \times \sigma^2(T_i), \tag{4}$$

where feature $x$ takes on values in $\{1, 2, \ldots, n\}$ and it splits the train set $T$ into subset $T_i$, $i \in \{1, 2, \ldots, n\}$. In this design, the success of approving a relevant feature relies on low entropy and low data variance. This is why the summation of the product of $H(x, T_i)$ and $\sigma^2(T_i)$ for all subsets becomes a better criterion for feature evaluation of multi-class problems.

In [37], no significant difference between the Gini index and information gain is found. The behavior of these two split functions is so similar that the probability in selecting the same splits is as high as 98%. Therefore, when numeric feature $x$ is tested for multi-class problems, the Gini index can be modified in a similar way as Shannon's entropy, i.e.,

$$\text{Enhanced\_Gini }(x, T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \text{Gini }(T_i) \times \sigma^2(T_i). \tag{5}$$

Similar to the design of classic information gain, $EH(x, T)$ is compared with the initial status before splitting. Such improvement is called *Aggregation Gain* and defined as $I(T) \times \sigma^2(T) - EH(x, T)$ in this paper, where $I(T)$ is the total entropy and $\sigma^2(T)$ is the total variance before splitting. $AG(x, T)$ is taken to represent the aggregation gain obtained by adopting feature $x$. Consequently, the higher the $AG(x, T)$, the more relevant feature $x$ is to the class. The effectiveness of information gain and aggregation gain will be discussed and analyzed in Section 5.

Both preprocessing of the raw data of feature values and feature evaluation constitute the first stage of our selection scheme. The goal of this stage is to investigate promptly all original features and rank them according to their discriminative power. The output of this stage is a small subset of candidate features for the subsequent classification. We remind that this stage can be accomplished within an execution complexity of $O(N)$.

## 4. Feature selection scheme

### 4.1. Variability analysis using statistical multivariate analysis

Redundancies among variables signify that their values correlate closely with each other. A variable becomes redundant because the data distribution and data variation of this variable has manifestations similar to those of other variables. Therefore, the class-discriminative power produced by two highly correlated features is similar to that derived from only one of them. Principal component analysis (PCA) is a multivariate statistical tool for analyzing data variability. It can reproduce the total system variability and achieves high reduction in dimensionality with usually lower noise than that of the original patterns. It is particularly helpful when simplifying a number of highly correlated variables into fewer independent components. The collection of independent variables is hard and impractical. Hence, it is difficult to prevent data redundancy when a set of variables is used. The goal of using PCA is to detect the degree of redundancy after a test variable is included into a subset of variables. In other words, if the test variable has a relative low dependency on a subset of variables, it can be selected as the next proper variable capable of the classification task. Inversely, the test variable with a high dependency on a subset of variables will be excluded from the classification task. The usages and mathematical background of PCA are given as follows.

In practice, the correlation matrix of the data is constructed and the eigenvectors on this matrix are computed. Correlation matrix can overcome the problem attributed to different measure scales. The eigenvectors that correspond to the largest eigenvalues (the principal components) can now be utilized to reconstruct a large fraction of the variance of the original data. The disadvantage of PCA is that the principle components are not easy to interpret. However, a simple classifier integrating multiple feature aspects without severely suffering from redundancy can be achieved by PCA.

Suppose there are $N$ features in a feature set $S$ (i.e., $S = \{f_1, f_2, \ldots, f_N\}$) and it consists of $m$ measurements on these $N$ features. A subset of $n$ features is taken from $S$ and the correlation matrix $\sum$ of these $n$ features is made for PCA. Let the eigenvalues of $\sum$ be $\lambda_1\ (S) \geqslant \lambda_2(S) \geqslant \cdots \geqslant \lambda_n(S) \geqslant 0$. The $i$-th component is the linear combination which is given by $y_i = a_{i1}\ f_1 + a_{i2}f_2 + \cdots + a_{in}f_n$, where $[a_{i1}, a_{i2}, \ldots, a_{in}]$ is the eigenvector corresponding to $\lambda_i(S)$, $i = 1, 2, \ldots, n$. Consequently, the proportion of total variance (i.e., explained variance) due to the $i$-th principal component is $\frac{\lambda_i(S)}{\sum_{j=1}^{n}\lambda_j(S)}$. The correlation coefficient between the component $y_i$ and the variable $f_k$ is $\rho_{i,k} = \frac{a_{ik}\sqrt{\lambda_i(S)}}{\sqrt{\sigma_k^2}}$, where $\sigma_k^2$ is the variance of $f_k$ and $i, k = 1, 2, \ldots, n$. Generally, most (for instance, more than 70%) of the total population variance can be featured to one or two components. In this paper, the eigenvalue corresponding to the first principal component is employed to approximate the data variability derived from a subset of features. In the case that a feature has redundant information for the existing selected features, it brings none or only little additional data variability to the existing system and this situation can be easily detected via PCA. The following theorem and corollary ensure that PCA can facilitate our selection scheme.

**Theorem 1.** *Let S be a system with N variables. If n variables are selected from S, $2 \leqslant n \leqslant N$, and PCA is applied to these variables such that n components $y_i$ $1 \leqslant i \leqslant n$, are extracted from the $n \times n$ correlation matrix $\Sigma$. The following three statements are valid for any n.*

(1) *For each $i \in [1, n]$, $\sigma^2(y_i) = \lambda_i$, where $\lambda_i$ is the* i-th *largest eigenvalue of $\Sigma$.*
(2) *Since $\sigma^2(y_1) \geqslant \sigma^2(y_2) \geqslant \cdots \geqslant \sigma^2(y_n)$, then $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$.*
(3) *$\sum_{i=1}^{n}\lambda_i = n$.*

**Proof.** Please refer to [15]. □

**Corollary 1.** *Let A be a subset comprising n variables of S and A′ is another subset such that $A' = A \cup v$, where v is a single variable selected from S–A. If $\lambda_i$, $1 \leqslant i \leqslant n$, is the eigenvalue of $\Sigma(A)$ and $\lambda_i'$, $1 \leqslant i \leqslant n+1$, is the eigenvalue of $\Sigma(A')$, then $\lambda_i' \geqslant \lambda_i$, for $1 \leqslant i \leqslant n$.*

**Proof.** If $v \neq 0$, i.e., $v$ is a non-zero vector, $\sum_{i=1}^{n+1}\lambda_i' = n+1 \geqslant \sum_{i=1}^{n}\lambda_i = n$ according to the third statement of the previous theorem. However, in the case that $v$ is a zero vector, $A'$ degenerates into $A$ so that $\lambda_i'$ is equal to $\lambda_i$ and $\lambda_{n+1}' = 0$, for $1 \leqslant i \leqslant n$. Therefore, it is concluded that the adding of a new variable can raise all original eigenvalues and generate a new additional eigenvalue. The sum of all increments plus the new additional eigenvalue is equal to 1. □

In our design, the change of the first eigenvalues ($\lambda_1$ and $\lambda_1'$) is investigated between two successive selection rounds. If a significant difference between $\lambda_1$ and $\lambda_1'$ is detected, the new additional feature has low redundancy for the current data system and it is identified as a serviceable feature. Otherwise, it has so little variability contribution to the current data system that it is classified as a useless feature. Users can designate the maximum number of rounds for the selection process. This number is utilized to limit the maximum number of selection rounds executed. If the variability analysis can acquire a satisfactory result before executing so many rounds, the selection process will cease and output the selected features.

### 4.2. Heuristic selection algorithm

For the completeness of our feature selection scheme, we present below the heuristic algorithm in detail. The notations used in the algorithm are as follows.

- $S$: The feature space comprising all original features whose size exceeds 2.
- $\varpi$: The designated maximum number of selection rounds.
- $A$, $A_1$, $A_2$: The sets comprising selected features.
- $\alpha$, $\beta$: The features selected in every selection round.
- $\lambda_1\ (S)$: The largest eigenvalue of the correlation matrix when applying PCA to $S$.
- $\psi$: The threshold between 0 and 1.

### 4.3. Heuristic selection

```
Input: feature set $S$ with training data
Output: set of selected features
1. Sort all features of $S$ in a decreasing order according to their
   individual $AG$s. If two or more features have the same
   evaluation, the feature with higher priority ranks first. The
   feature with the highest $AG$ is denoted as $f_1$.
2. $A \leftarrow \{f_1\}$.
3. While $|A| \leqslant \varpi$
4.     Select the next feature $\alpha$ in $S - A$ which maximizes
       $\lambda_1(A \cup \alpha)$.
5.     $A_1 \leftarrow A \cup \{\alpha\}$.
6.     Similarly, select the next feature $\beta$ in $S - A_1$ which
       maximizes $\lambda_1(A_1 \cup \beta)$.
7.     $A_2 \leftarrow A_1 \cup \{\beta\}$.
8.     If $\frac{\lambda_1(A_2)-\lambda_1(A_1)}{\lambda_1(S)} \geqslant \psi$, then $A \leftarrow A_2$.
9.     Else $A \leftarrow A_1$ and break the while loop.
10. End While
11. Return $A$
```

Initially, the feature with the best $AG$ evaluation serves as the first basis for the next round of selection. As line 3 shows, the algorithm may keep selecting before reaching the maximum number of selection rounds $\varpi$. A feasible $\varpi$ of $\left|\sqrt{N}\right| + 1$ is suggested according to the experience rule of statistical sampling, where $N$ is the size of the feature space $S$. By means of variability analyses, lines 4–7 select two features that have better complementary variability effect on set $A$. Line 8 measures the incremental variability ratio gained from the new additional feature. The measurement exceeding the threshold $\psi$ means that the enhancement is so significant that the feature should be collected. Under such situation, another round of selection will be triggered. Oppositely, insignificant enhancement will cause the selection process to be terminated as shown in line 9. Line 11 outputs the final results. The proceeding of this heuristic selection is controlled by $\varpi$ and $\psi$. Lower $\varpi$ and higher $\psi$ will yield a compact feature subset while higher $\varpi$ and lower $\psi$ will output a plentiful subset.

Fig. 2 shows the entire scheme of our feature selection. It is assumed that the readers are familiar with FCM cluster analysis validated by the Huang-index as mentioned in Section 2. Four steps including first cluster processing, feature evaluation, feature selection, and second cluster processing are respectively highlighted as follows.

I. *First cluster processing*: before evaluating the relevance of features with respect to the target variable, raw data with real-valued and ordinal features are first preprocessed by cluster analysis. Nominal data are exempted from this preprocessing. Our clustering method can group raw data into a non-uniform categorization and assign every cluster a sequencing number. In this step, all features are equally respected without any selective priority. The output of this step is the discrete values corresponding to the processed features. Assume there are $M$ instances in the dataset. The data complexity involved in this step is $O(\widetilde{N} \times M)$, where $\widetilde{N}$ is the number of real-valued and ordinal features and $\widetilde{N} \leqslant N$. Regarding computational complexity, $O(\widetilde{N})$ is required in this step.

II. *Feature evaluation*: all input features are evaluated by our $AG$ criterion. All data used in this individual evaluation strategy are only involved once. The data complexity involved in this step is $O(N \times M)$. This step needs at most an execution complexity of $O(N)$.

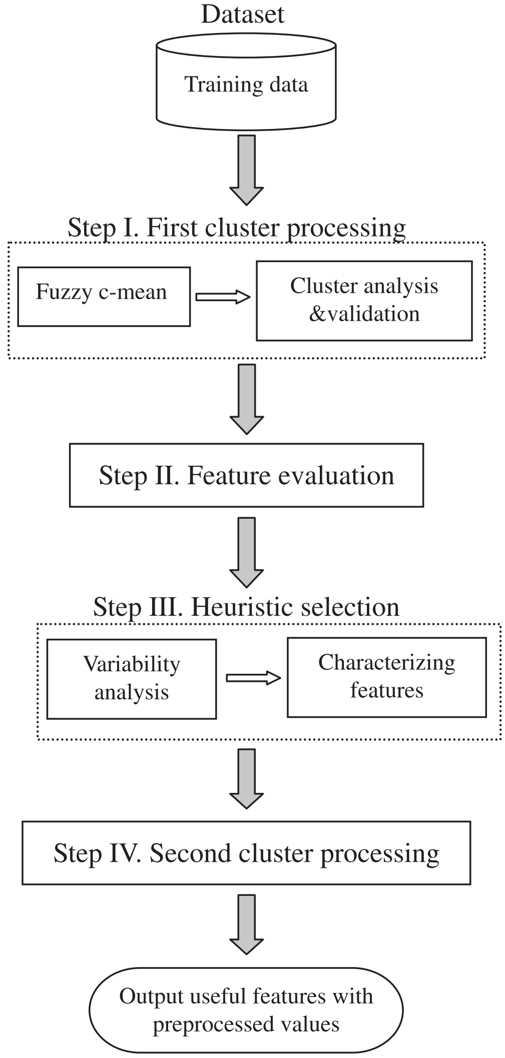

Dataset
Training data
Step I. First cluster processing
Fuzzy c-mean
Cluster analysis &validation
Step II. Feature evaluation
Step III. Heuristic selection
Variability analysis
Characterizing features
Step IV. Second cluster processing
Output useful features with preprocessed values

**Fig. 2.** Feature selection scheme.

III. *Heuristic selection*: the computation of correlation matrices in variability analyses requires higher computational costs. However, only a few features are involved in this step. A small number of $\left|\sqrt{N}\right| + 1$ features are suggested to be sufficient for this step. Therefore, it needs an execution complexity of $O\left(\left(\sqrt{N}+1\right)^2\right) = O(N)$ and the involved data complexity is $O(N^{1/2} \times M)$. How the number of selected features affects the classification performance will be discussed in the next section.

IV. *Second cluster processing*: only the features selected in Step III are processed in this step. The raw data corresponding to these features are recalled and processed again by our cluster analysis. The second cluster processing can result in a higher number of clusters than that of the first cluster processing. Since only a quite small number of features ($\leqslant N^{1/2}$) are processed by the second cluster analysis, the execution complexity and data complexity involved in this step are respectively less than $O(N^{1/2})$ and $O(N^{1/2} \times M)$.

The overall data complexity involved in the computation of our heuristic selection scheme is $O(N \times M)$, which is comparable with that of many other feature selection algorithms. On the other hand, the total computation complexity or our scheme is $O(N)$ and this

**Table 1**
Abstract of five datasets.

| Datasets | *Segment* | *Places* | *Wine* | *Svmguige4* | *Satimage* |
|---|---|---|---|---|---|
| # of instances | 2310 | 987 | 178 | 612 | 4435 |
| # of features | 19 | 9 | 13 | 10 | 36 |
| # of classes | 7 | 4 | 3 | 6 | 6 |

retains the same economic level as those algorithms classified into the *filter* category.

## 5. Experimental results and analyses

### 5.1. Dataset acquisition

The five multi-class datasets used in this paper are named as *segment*, *places*, *wine*, *svmguide*4, and *satimage*. They were respectively downloaded from Statlog [42], UCI [45], or StatLib [6,41] and their application domains cover image analysis, chemical analysis, and traffic light signal. Table 1 depicts the information of these datasets. The number of features excluding the target class varies from 9 to 36 and the number of target classes varies from 3 to 7. Dataset *segment* has the highest number of classes while dataset *satimage* has the largest instance size and number of features among all. The raw data values corresponding to all features are either real or integer. Five datasets of preprocessed instances are saved into five individual files so that they can be reused for competitive experiments. The implementation of the proposed scheme from Step I to Step IV was executed in *C* and *Matlab* programming languages performed on a workstation with an Intel Core 2 dual 2.4 GHz processor.

### 5.2. Relevant feature authentication

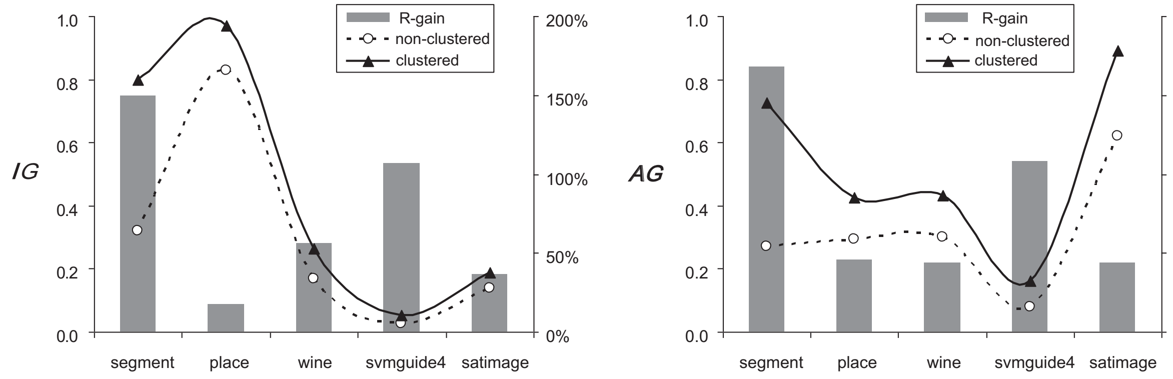
The first experiments compare the discriminative effects between criteria *IG* and *AG*. Real data are processed with discretization while integer data retain their numeric order. All data are saved into five individual files so that they can be reused for clustering process and feature evaluation. We compare the discriminative power between the raw data processed by discretization and those involved in cluster analyses. Fig. 3a and b show respectively the average *IG*s and *AG*s measured from all features in the five datasets. Note that there are two vertical axes in the figures. Solid lines connected by triangle legends display the evaluations of raw data processed by cluster analyses while the dashed lines connected by circle legends show the evaluations of raw data processed by discretization. These evaluations are referred to as graduations on the left axes. The criterion which generates higher evaluation values helps distinguish explicitly features on a more apparent condition. As can be seen, clustered data have higher *IG*s and *AG*s than non-clustered data, revealing that cluster analyses can boost the discriminative effects of raw data. In addition, the ratios of difference gain (denoted as *R*-gain) are displayed as the gray bars in the figures, where *R*-gain is defined as $(IG^{clustered}\text{-}IG^{non-clustered})/IG^{non-clustered}$ or $(AG^{clustered}\text{-}AG^{non-clustered})/AG^{non-clustered}$. These bars are referred to as graduations on the right axes. We note that, in Fig. 3b, the *AG*s of the clustered data have at least 45% growth as compared with those of the non-clustered data.

The features with better evaluations are selected for further investigation. The amount of selected features is designated as $\left|\sqrt{N}\right| + 1$, where $N$ is the number of features for each dataset. As can be seen, Fig. 4a has performance similar to that of Fig. 3a. However, the *R*-gains in Fig. 4b have significant growth as compared with those in Fig. 3b. These results signify that the difference in discriminative effects of good features between clustered and non-clustered data is readily identified by *AG* criteria.

Fig. 5a and b outline the *R*-gains for all features and the better $\left|\sqrt{N}\right| + 1$ features, respectively. As seen in Fig. 5a, the difference between *AG*s and *IG*s is so negligible that it is difficult to choose between the two criteria. However, as shown in Fig. 5b, the *AG*s of the good features outperform the *IG*s of the good features for all datasets except for the dataset *wine*. Two factors dominate this phenomenon. The first is data of their own distribution and variation, i.e., whether data have consistent or inconsistent significance. The second is the degree of complexity of the target classification variable. *AG* favors multi-class cases and this is why dataset *wine* does not acquire much improvement when evaluating the discrimination effects of features.

### 5.3. Performance of selected features

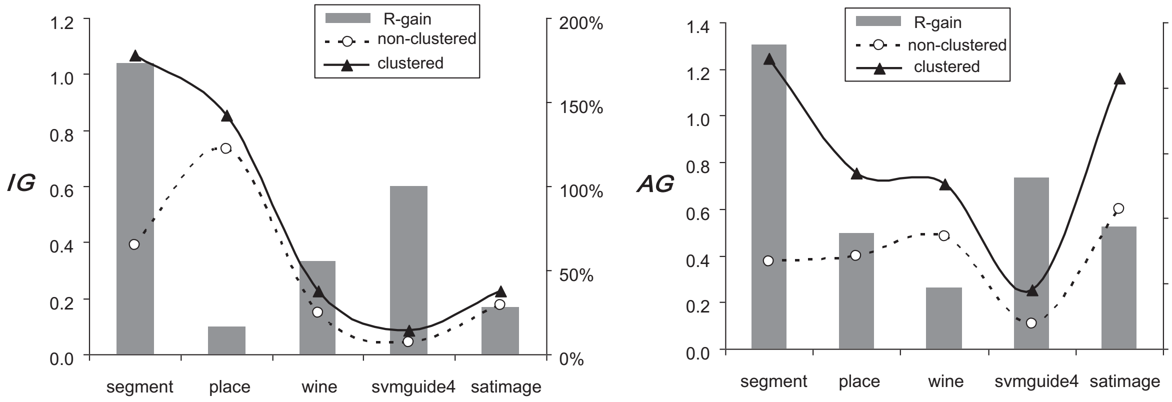
To verify the effectiveness of the selected features, four classification methods including NaiveBayes, C4.5, simple CART, and SVM, which were selected from the 10 most influential algorithms [52] were tested. Different feature subsets were generated according to three strategies. The first strategy only adopts the feature evaluation criterion of *information gain*, which is the classic relevance analysis and marked as “*R*” in the following figures. In addition to the classic relevance analysis, the second strategy also applies the variability analysis via PCA to reduce the redundancy among the selected features, which is denoted as “RV”in the following figures. The third strategy adopts the relevance analysis of *AG* criterion, the variability analysis of PCA, and the cluster analysis



(a) Average *IG*s

(b) Average *AG*s

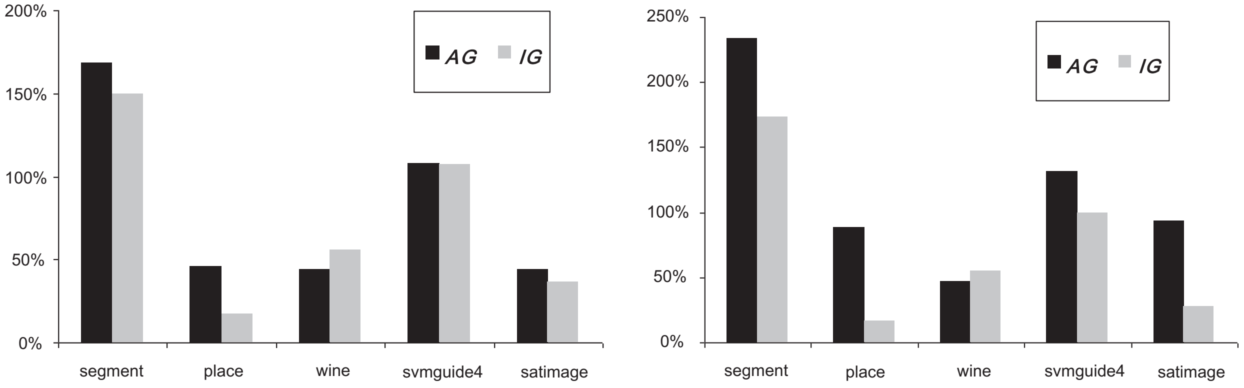
**Fig. 3.** Evaluations for all features in five datasets.



(a) Average *IG*s

(b) Average *AG*s

**Fig. 4.** Evaluations for the better $\left|\sqrt{N}\right| + 1$ features in five datasets.



(a) All features

(b) Better $\left\lceil\sqrt{N}\right\rceil$+ 1 features

**Fig. 5.** Ratios of difference gain of average *AG*s and *IG*s in five datasets.

validated by the Huang-index, which is marked as "RVC"in the following figures. Since the effectiveness of cluster analyses is addressed in Section 5.2, the performance gained from relevance analysis plus cluster analysis (i.e., "RC") was not included in this experiment. To investigate the classification performance affected by number of features as well, three feature quantities of $\left\lceil\sqrt{N/2}\right\rceil + 1$, $\left|\sqrt{N}\right| + 1$, and $\left|\sqrt{2N}\right| + 1$ were tested. For simplicity, these three quantities are respectively denoted as $N^-$, $N$, and $N^+$ in the following figures.

Figs. 6a–d respectively show the accuracy performance of four classifiers in different cases. All experimental results in this study were assessed using 10-fold cross-validation. The features selected by RVC generally outperform those by *R* or RV. Dataset *svmguide*4 has the most significant improvement for all four classifiers. The three feature quantities used in *svmguide*4 were $N^- = 4$, $N = 5$, and $N^+ = 6$. The best performance all happened at $N = 5$. As to datasets *segment*, *places*, and *satimage*, the benefits brought by cluster analyses seem to be larger than those brought by variability analyses. In other words, the performance achieved by *R* and RV is similar in some cases (i.e., *places* and *satimage* in C4.5, simple CART, and SVM). The effect of variability analysis is slightly apparent only when *segment* and *svmguide*4 are used. In view of all the experimental results shown in Fig. 6, variability analysis can enhance the effect of cluster analyses in most cases except for *wine*. In dataset *wine*, similar performance for the three strategies reveals that variability and cluster analyses do not contribute much to the classification tasks because the selected features are inherently uncorrelated and the raw data are invariant under the processing of cluster analyses. With respect to the factor of feature quantity, both *segment* and *satimage* attain improvement when more features are imposed on the cases using *R* and RV. However, RVC can achieve high accuracy before involving the high feature quantity of $N^+$.

The discrimination capability of classifiers is typically measured by the ROC area. The ROC area is directly represented by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate). It is a comparison of two operating characteristics (TPR and FPR) as the criterion changes and therefore measures the discrimination capability of the classifier. The closer the curve to the upper left-hand corner of the graph, the greater the area and the higher the discrimination capability is. The range of the ROC area is 0–1. An area of 1 represents a perfect test and an area of 0.5 represents a worthless test. A rough guide [57] for classifying the discrimination capability of a classifier is as follows.

- *Excellent*: 0.9–1.
- *Good*: 0.80–0.9.
- *Fair*: 0.7–0.8
- *Poor*: 0.6–0.7.
- *Failed*: 0.5–0.6.

As shown in Fig. 7, the features selected by RVC have successfully raised discrimination capability of classifiers to "Good" or

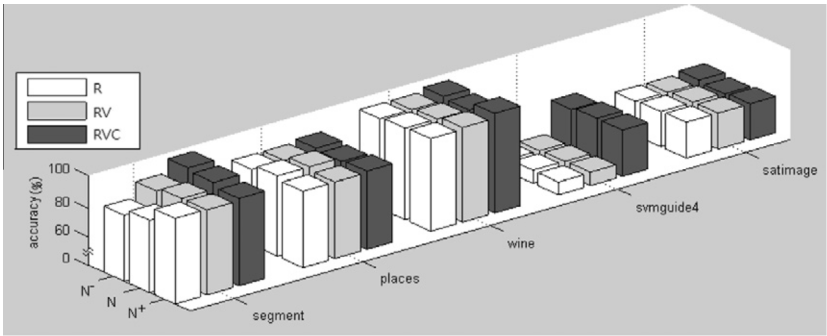


**Fig. 6a.** NaiveBayes.

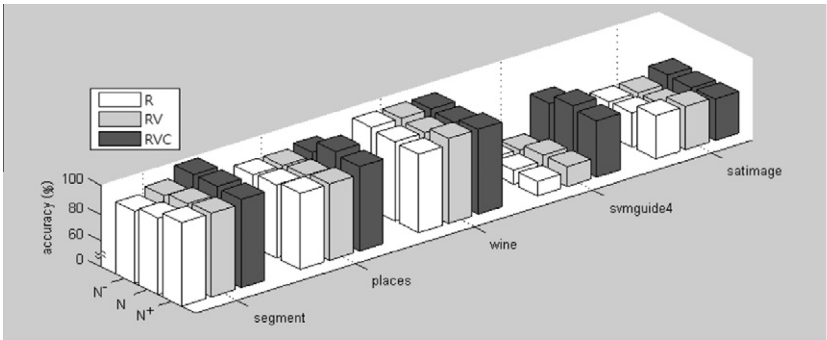


**Fig. 6b.** C4.5.

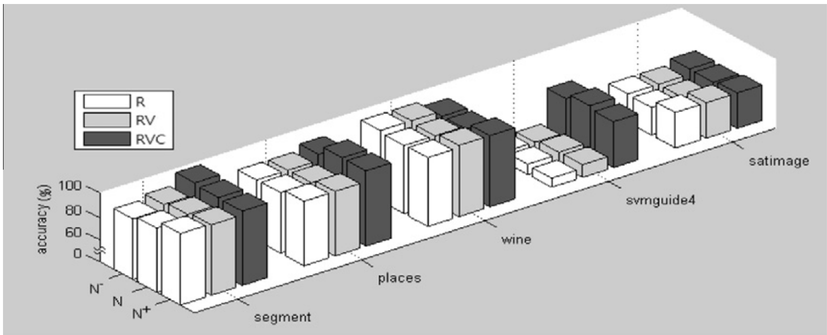


**Fig. 6c.** Simple CART.

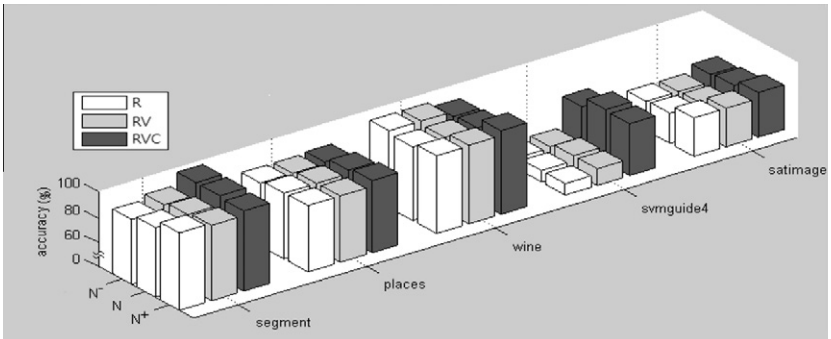


**Fig. 6d.** SVM.

even higher levels. In view of the performance obtained by *places* and *svmguide*4, the effect of cluster analyses and variability analyses is verified once again. On the other hand, selected feature quantity has no remarkable influence on the experimental results. This phenomenon is consistent with the well-known fact that only a limited number of relevant features are sufficient to complete the entire classification task.

Table 2 lists the accuracies and ROCs (in brackets) achieved by RVC for the four classifiers and five datasets with the best results being highlighted. Our feature selection scheme does not seem to

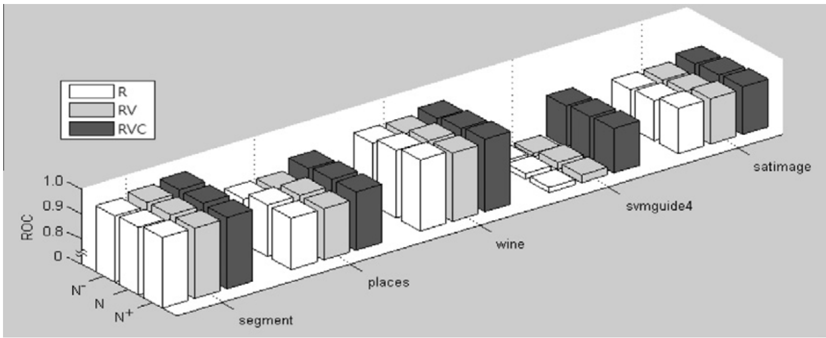


**Fig. 7a.** NaiveBayes.

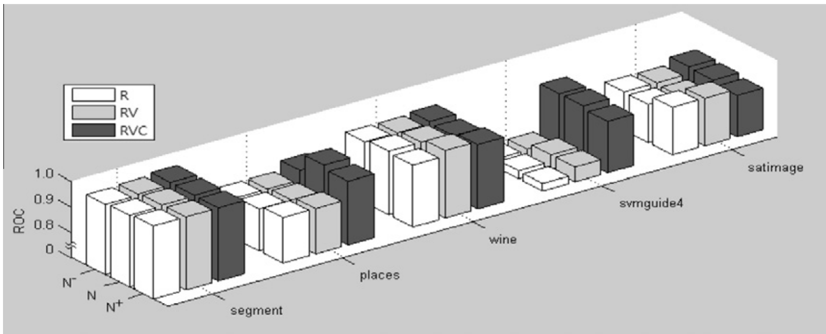


**Fig. 7b.** C4.5.

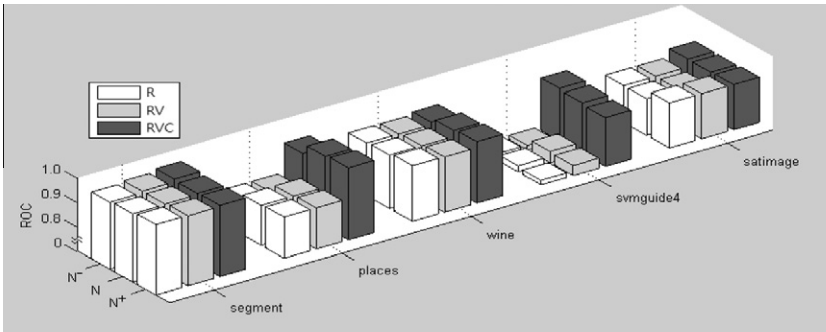


**Fig. 7c.** Simple CART.

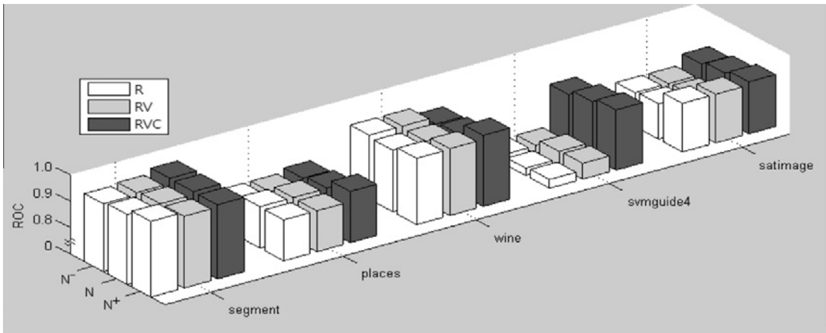


**Fig. 7d.** SVM.

favor any classifier since the difference is insignificant. The last column of Table 2 shows the total average results for the five datasets. It is worth noting that our scheme can supply high discrimination quality to four classifiers since all ROCs exceed 0.93.

Last but not least, to prove the usefulness of our RVC method, three well-known conventional schemes of feature selection including *Information Gain* (IG), *Chi-Squared* ($\chi^2$) [27], and *Correlation-based Feature Selection* (CFS) [16,17] methods are employed in

**Table 2**
Accuracies (%) and ROCs of five datasets for four classifiers.

| | *Segment* | *Places* | *Wine* | *Svmguide4* | *Satimage* | Average |
|---|---|---|---|---|---|---|
| Naïve Bayes | 91[0.988] | 83[0.928] | 93[0.978] | 65[0.871] | 80[0.914] | 82.4[0.936] |
| C4.5 | 94[0.989] | 90[0.936] | 90[0.945] | 74[0.910] | 85[0.932] | 86.6[0.942] |
| Simple CART | 95[0.990] | 92[0.977] | 89[0.940] | 73[0.907] | 85[0.923] | 86.8[0.947] |
| SVM | 94[0.983] | 86[0.867] | 91[0.936] | 73[0.910] | 87[0.952] | 86.2[0.930] |

**Table 3a**
Performance of four feature selection schemes for Naïve Bayes.

| | RVC | IG | $\chi^2$ | CFS |
|---|---|---|---|---|
| *Segment* | 91[0.988] | 69[0.914] | 75[0.896] | 92[0.912] |
| *Places* | 83[0.928] | 56[0.836] | 74[0.902] | 80[0.825] |
| *Wine* | 93[0.978] | 62[0.877] | 73[0.889] | 89[0.916] |
| *Svmguide4* | 65[0.871] | 42[0.691] | 53[0.810] | 68[0.796] |
| *Satimage* | 80[0.914] | 74[0.938] | 73[0.871] | 78[0.854] |
| Average | 82.4[0.936] | 60.6[0.851] | 69.6[0.874] | 81.4[0.861] |

**Table 3b**
Performance of four feature selection schemes for C4.5.

| | RVC | IG | $\chi^2$ | CFS |
|---|---|---|---|---|
| *Segment* | 94[0.989] | 71[0.921] | 75[0.924] | 93[0.915] |
| *Places* | 90[0.936] | 60[0.856] | 74[0.882] | 88[0.903] |
| *Wine* | 90[0.945] | 61[0.867] | 65[0.873] | 90[0.904] |
| *Svmguide4* | 74[0.910] | 43[0.691] | 55[0.792] | 70[0.784] |
| *Satimage* | 85[0.932] | 81[0.936] | 82[0.934] | 86[0.853] |
| Average | 86.6[0.942] | 63.2[0.854] | 70.2[0.881] | 85.4[0.872] |

**Table 3c**
Performance of four feature selection schemes for simple CART.

| | RVC | IG | $\chi^2$ | CFS |
|---|---|---|---|---|
| *Segment* | 95[0.990] | 70[0.918] | 68[0.843] | 95[0.918] |
| *Places* | 92[0.977] | 62[0.824] | 70[0.903] | 93[0.905] |
| *Wine* | 89[0.940] | 60[0.845] | 63[0.867] | 89[0.902] |
| *Svmguide4* | 73[0.907] | 42[0.697] | 43[0.703] | 75[0.813] |
| *Satimage* | 85[0.923] | 81[0.828] | 75[0.831] | 83[0.838] |
| Average | 86.8[0.947] | 63.0[0.822] | 63.8[0.829] | 87.0[0.875] |

**Table 3d**
Performance of four feature selection schemes for SVM.

| | RVC | IG | $\chi^2$ | CFS |
|---|---|---|---|---|
| *Segment* | 94[0.983] | 70[0.903] | 69[0.871] | 93[0.915] |
| *Places* | 86[0.867] | 59[0.845] | 72[0.895] | 90[0.882] |
| *Wine* | 91[0.936] | 61[0.881] | 73[0.822] | 90[0.904] |
| *Svmguide4* | 73[0.910] | 43[0.670] | 55[0.830] | 71[0.741] |
| *Satimage* | 87[0.952] | 80[0.926] | 76[0.845] | 88[0.862] |
| Average | 86.2[0.930] | 62.6[0.845] | 69[0.853] | 86.4[0.861] |

the final experiment. The reason why these methods are chosen is that they are all built on the top of the entropy criterion. The IG and $\chi^2$ methods evaluate features individually by measuring their information amount and chi-squared statistic with respect to the classes. Rather than merely scoring individual features, RVC and CFS methods analyze (or evaluate) the worth of subsets of features. The IG, $\chi^2$, and CFS methods do not handle the similarity or proximity of feature values before processing feature evaluation. In addition, the heuristic in RVC is different from that in CFS. We remind that the data complexity involved in the execution of RVC heuristic is $O(N \times M)$, where $N$ is the initial number of features and $M$ is the instance number in the original dataset. However, the CFS method requires $O((N^2 - N)/2) \times M)$ operations [16] for computing the pairwise feature correlation matrix. Hence, the computational cost of RVC is more economic than that of CFS. The 10-fold cross-validation classification accuracies (%) and ROCs for four classifiers are listed in Tables 3a–d and the experimental results are averaged and listed in the last rows of these tables. The RVC method outperforms the IG and $\chi^2$ methods simultaneously in the aspects of classification accuracy and discrimination power. Approximately, 30% accuracy improvement and 10% discrimination improvement are achieved in these datasets. As stated in [36], the *m* best features are not the best *m* features. So, combinations of individually good features do not necessarily lead to good classification performance. These improvements of RVC rely on the cooperation of feature cluster analysis, the *AG* criterion, and variability analysis. When comparing with the CFS method, the discrimination power derived from RVC outperforms that derived from CFS. The average gain is about 8%. As to classification accuracy, although RVC does not significantly outperform CFS, it is very competitive to CFS in these datasets.

## 6. Conclusions

Multi-class classification problems incur more intricate conditions than binary classification problems. These conditions have pushed classification problems to the further boundary where more techniques and technologies are drawn into the data mining community. The feature selection scheme proposed in this paper has three main contributions. First, it is the pioneering attempt to promote the discriminative power of raw data by means of cluster analyses. Although both FCM and Huang-index method yield superior effects when evaluating features, many other clustering algorithms can realize the same objective as well. Second, more precise evaluation criteria have been proposed for multi-class problems. Only limited computational costs are required in this design. The whole evaluation complexity remains at the same level. Third, the compact subset comprising discriminative features is acquired by the heuristic selection scheme. Cluster, relevance, and variability analyses are successfully integrated into an efficient and effective framework. Future researches should be directed to the following aspects. The first is to apply different cluster analyses to different features and use a distinct number of clusters to assess every single features. We are motivated to raise the discrimination power of features to an even higher level; and success in achieving such can reduce the number of features involved in the classification model. In addition, speeding multivariate statistical analyses could be another issue when a large number of selected features is unavoidable. Finally, the creation of a new classifier which can fully advance the superiority of the new feature selection scheme over many traditional methods, is needed and of help.

## References

[1] O. Aran, L. Akarun, A multi-class classification strategy for Fisher scores: application to signer independent sign language recognition, Pattern Recognition. 43 (5) (2010) 1776–1788.

[2] J.C. Bezdek, Cluster validity with fuzzy sets, Journal of Cybernetics 3 (3) (1973) 58–73.

[3] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms (Advanced Applications in Pattern Recognition) Publisher: Plenum Press - New York, 1981.

[4] J.C.F. Caballero, F.J. Martínez, C. Hervás, P.A. Gutiérrez, Sensitivity versus accuracy in multiclass problems using memetic pareto evolutionary neural networks, IEEE Transactions on Neural Networks 21 (5) (2010) 750–770.

[5] M.Á. Carreira-Perpiñán, Continuous Latent Variable Models for Dimensionality Reduction and Sequential data Reconstruction. PhD dissertation, Dept. of Computer Science, Univ. of Sheffield, UK, 2001.

[6] C.-C. Chang, C.-J. Lin, Software Available: A Library for Support Vector Machines (LIBSVM), 2001. <http://www.csie.ntu.edu.tw/∼cjlin/libsvm>.

[7] J. Chen, C. Wang, R. Wang, Adaptive binary tree for fast SVM multiclass classification, Neurocomputing 72 (13–15) (2009) 3370–3375.

[8] M. Dash, K. Choi, P. Scheuermann, H. Liu, Feature selection for clustering – a filter solution, in: Proc. Second Int'l Conf., 2002, pp. 115–122.

[9] F. Fleuret, Fast binary feature selection with conditional mutual information, Journal of Machine Learning Research 5 (2004) 1531–1555.

[10] I.K. Fodor, A Survey of Dimension Reduction Techniques, Technical Report UCRL-ID-148494, Lawrence Livermore National Laboratory, Center for Applied Scientific Computing, 2002.

[11] J.F. Gantz, A Forecast of Worldwide Information Growth Through, March 2007, 2010. <http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf>.

[12] A. Gupta, K.G. Mehrotra, C. Mohan, A clustering-based discretization for supervised learning, Statistics and Probability Letters 80 (9–10) (2010) 816–824.

[13] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.

[14] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning 46 (1–3) (2002) 389–422.

[15] J.F. Hair, W.C. Black, B.J. Babin, R.E. Anderson, R.L. Tatham, Multivariate Data Analysis. Pearson Prentice Hall, NY, USA, 2006.

[16] M.A. Hall, Correlation-based Feature Selection Machine Learning, Ph.D. Thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1998.

[17] M.A. Hall, L.A. Smith, Feature subset selection: a correlation based filter approach, in: Proceeding of the Fourth International Conference on Neural Information Processing and Intelligent Information Systems, 1997, pp. 855–858.

[18] D. Huang, T.W.S. Chow, Effective feature selection scheme using mutual information, Neurocomputing 63 (2005) 325–343.

[19] J. Huang, Y. Cai, X. Xu, A hybrid genetic algorithm for feature selection wrapper based on mutual information, Pattern Recognition Letters 28 (13) (2007) 1825–1844.

[20] K.-Y. Huang, Applications of an enhanced cluster validity index method based on the fuzzy *c*-means and rough set theories to partition and classification, Expert Systems with Applications 37 (12) (2010) 8757–8769.

[21] K.-Y. Huang, An enhanced classification method comprising a genetic algorithm, rough set theory and a modified PBMF-index function, Applied Soft Computing 12 (1) (2012) 46–63.

[22] M.M. Kabir, M.M. Islam, K. Murase, A new wrapper feature selection approach using neural network, Neurocomputing 73 (16–18) (2010) 3273–3283.

[23] D.-W. Kim, K.H. Lee, D. Lee, On cluster validity index for estimation of the optimal number of fuzzy clusters, Pattern Recognition 37 (10) (2004) 2009–2025.

[24] H.-M. Lee, C.-M. Chen, J.-M. Chen, Y.-L. Jou, An efficient fuzzy classifier with feature selection based on fuzzy entropy, IEEE Transactions on Systems, Man, and Cybernetics 31 (3) (2001) 426–432.

[25] R. Li, J. Lu, Y. Zhang, T. Zhao, Dynamic Adaboost learning with feature selection based on parallel genetic algorithm for image annotation, Knowledge-Based Systems 23 (3) (2010) 195–201.

[26] M. Lindenbaum, S. Markovitch, D. Rusakov, Selective sampling for nearest neighbor classifiers, Machine Learning 54 (2) (2004) 125–152.

[27] H. Liu, R. Setiono, Chi2: feature selection and discretization of numeric attributes, in: Proc. IEEE 7th International Conference on Tools with Artificial Intelligence, 1995, pp. 338–391.

[28] S. Maldonado, R. Weber, J. Basak, Simultaneous feature selection and classification using kernel-penalized support vector machines, Information Sciences 181 (1) (2011) 115–128.

[29] M. Medvedovic, K.Y. Yeung, R.E. Bumgarner, Bayesian mixture model based clustering of replicated microarray data, Bioinformatics 20 (8) (2004) 1222–1232.

[30] P. Mitra, C.A. Murthy, S.K. Pal, Unsupervised feature selection using feature similarity, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (3) (2002) 301–312.

[31] M.A. Muharram, G.D. Smith, Evolutionary feature construction using information gain and gini index, Lecture Notes in Computer Science 3003 (2004) 379–388.

[32] R. Nock, F. Nielsen, On weighting clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (8) (2006) 1223–1235.

[33] G. Ou, Y.L. Murphey, Multi-class pattern classification using neural networks, Pattern Recognition 40 (1) (2007) 4–18.

[34] M.K. Pakhira, S. Bandyopadhyay, U. Maulik, A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification, Fuzzy Sets and Systems 155 (2) (2005) 191–214.

[35] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishers, Dordrecht, 1991.

[36] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1226–1238.

[37] L.E. Raileanu, K. Stoffel, Theoretical comparison between the gini index and information gain criteria, Annals of Mathematics and Artificial Intelligence 41 (1) (2004) 77–93.

[38] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.

[39] S. Saha, S. Bandyopadhyay, Application of a new symmetry-based cluster validity index for satellite image segmentation, IEEE Geoscience and Remote Sensing Letters 5 (2) (2008) 166–170.

[40] A.I. Schein, L.H. Ungar, Active learning for logistic regression: an evaluation, Machine Learning 68 (3) (2007) 235–265.

[41] StatLib Databsets Archive. <http://lib.stat.cmu.edu/datasets/>.

[42] Statlog Datasets. <http://www.is.umk.pl/projects/datasets.html>.

[43] C. -T Su, Y.-H. Hsiao, Multiclass MTS for simultaneous feature selection and classification, IEEE Transactions on Knowledge and Data Engineering 21 (2) (2009) 192–205.

[44] C.-F. Tsai, Y.-C. Hsiao, Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches, Decision Support Systems 50 (1) (2010) 258–269.

[45] UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml/>.

[46] C.-M. Wang, Y.-F. Huang, Evolutionary-based feature selection approaches with new criteria for data mining: a case study of credit approval data, Expert Systems with Applications 36 (3) (2009) 5900–5908.

[47] H.-L. Wei, S.A. Billings, Feature subset selection and ranking for data dimensionality reduction, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (1) (2007) 162–166.

[48] J. Weston, A. Elisseeff, B. Schoelkopf, M. Tipping, Use of the zero-norm with linear models and kernel methods, Journal of Machine Learning Research 3 (2003) 1439–1461.

[49] I. Witten, E. Frank, A Practical Machine Learning Tool with Java Implementation, Morgan Kaufmann Publishers, San Francisco, California, 2000.

[50] K.-L. Wu, M.-S. Yang, A cluster validity index for fuzzy clustering, Pattern Recognition Letters 26 (9) (2005) 1275–1291.

[51] S. Wu, T.W.S. Chow, Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density, Pattern Recognition 37 (2) (2004) 175–188.

[52] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, et al., Top 10 algorithms in data mining, Knowledge and Information Systems 14 (1) (2008) 1–37.

[53] L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, Journal of Machine Learning Research 5 (2004) 1205–1224.

[54] X.D. Zhang, M. Ferrer, A.S. Espeseth, S.D. Marine, E.M. Stec, M.A. Crackower, D.J. Holder, J.F. Heyse, B. Strulovici, The use of strictly standardized mean difference for hit selection in primary RNA interference high-throughput screening experiments, Journal of Biomolecular Screening 12 (4) (2007) 497–509.

[55] X.D. Zhang, R. Lacson, Y. Ruojing, S.D. Marine, A. Mccampbell, D.M. Toolan, T.R. Hare, J. Kajdas, J.P. Berger, D.J. Holder, J.F. Heyse, M. Ferrer, The use of SSMD-based false discovery and false nondiscovery rates in genome-scale RNAi screens, Journal of Biomolecular Screening 15 (9) (2010) 1113–1123.

[56] Y. Zhang, W. Wang, X. Zhang, Y. Li, A cluster validity index for fuzzy clustering, Information Sciences 178 (4) (2008) 1205–1218.

[57] http://gim.unmc.edu/dxtests/roc3.htm.