Neurocomputing 225 (2017) 11–22

Contents lists available at ScienceDirect

Neurocomputing

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

Feature selection based on quality of information

Jinghua Liu[a,b], Yaojin Lin[b,*], Menglei Lin[b], Shunxiang Wu[a], Jia Zhang[b]

[a] Department of Automation, Xiamen University, Xiamen 361000, PR China
[b] School of Computer Science, Minnan Normal University, Zhangzhou 363000, PR China

ARTICLE INFO

Communicated by Gang Zeng

Keywords:
Feature selection
Information entropy
Maximum-nearest-neighbor
Quality of information

ABSTRACT

Feature selection as one of the key problems of data preprocessing is a hot research topic in pattern recognition, machine learning, and data mining. Evaluating the relevance between features based on information theory is a popular and effective method. However, very little research pays attention to the distinguishing ability of feature, i.e., the degree of a feature distinguishes a given sample with other samples. In this paper, we propose a new feature selection method based on the distinguishing ability of feature. First, we define the concept of maximum-nearest-neighbor, and use this concept to discriminate the nearest neighbors of samples. Then, we present a new measure method for evaluating the quality of feature. Finally, the proposed algorithm is tested on benchmark datasets. Experimental results show that the proposed algorithm can effectively select a discriminative feature subset, and performs as well as or better than other popular feature selection algorithms.

1. Introduction

Feature selection as an important data preprocessing technique is widely applied in pattern recognition and machine learning. With the development of information storage, input data have a large number of features, which may include a mass of irrelevant and/or redundant features. Unnecessary features often result in "dimensionality curse", the low efficiency of learning algorithm, and the problem of over-fitting [5,13,25,40]. Therefore, it is an extremely important step to select the relevant and necessary features for a given learning task.

Feature selection is a process of selecting an optimal feature subset from the raw feature space according to a specific evaluation criterion. An optimal evaluation criterion can make the dimension of features much smaller compared to the number of original feature space, but as much information as possible is retained in the selected feature subset [2,7,16,32,38,43]. According to the evaluation criteria, existing feature selection methods can be classified into three models [10,13,28,49]: the wrapper model, the embedded model, and the filter model. In the wrapper model, the performance of feature selection depends on classifier directly, such as IWSS [6], GA/FDA [11], and RFE [14]. The embedded model needs a specific learning algorithm before conducting feature selection in the process of training, like inherently binary decision tree classifiers [8]. The filter model is usually to maximize the evaluation function for getting an optimal feature subset through a search strategy, and the evaluation function mainly includes distance metrics [20,21,50], dependency metrics [1,23,47], consistency [2,44], and information metrics [16,24,35,41].

Compared to the wrapper and the embedded models, the filter model not involves any learning machine in the process of feature selection. Therefore, the present studies pay more attention on the filter method, such as ReliefF [21], CFS [15], FCBF [42], and mRMR [32]. ReliefF estimates the importance of feature according to the feature's weight by searching the nearest neighbors from the same and different classes of a given sample. CFS (Correlation-based Feature Selection) takes account both the correlation of feature and the effectiveness of feature to predict the class label. FCBF (Fast Correlation-based Filter) considers both relevance and redundancy with two steps. It firstly employs the symmetrical uncertainty(SU) to describe the relevance between the candidate feature f and the class label d, then sorts features in descending order. Finally, it removes redundant features from the order list. Similarly, mRMR (minimum-Redundancy-Maximum-Relevancy) generates feature subset by selecting minimal redundancy and maximal relevance feature. Following these methods, Liu et al. [27] regarded feature selection as a feature clustering procedure, and used mutual information as a measure criterion for the between-cluster distance. Mutual information as one of filter methods is used to measure the quality of feature by assessing the correlation and redundancy of feature, and is also used to search the optimal feature for splitting information in ID3 and C4.5 [33,34]. Among these methods, information measure is effectively discussed about the relevance between condition features and decision feature.

Moreover, different search strategies have been designed to reduce the probability of missing a good feature subset. From the view of search strategy, feature selection can be grouped into three types [36]:

* Corresponding author.
E-mail addresses: zzliujinghua@163.com (J. Liu), yjlin@mnnu.edu.cn (Y. Lin), menglei36@126.com (M. Lin), sxwu@xmu.edu.cn (S. Wu), zhangjia_gl@163.com (J. Zhang).

http://dx.doi.org/10.1016/j.neucom.2016.11.001
Received 15 January 2015; Received in revised form 8 June 2016; Accepted 1 November 2016
Available online 09 November 2016

complete search, stochastic search, and heuristic search. Complete search can find a best feature subset. However, it is difficult to deal with the medium and/or big sized feature sets because the high computational complexity. Stochastic search mainly balances optimality and manageability for selecting feature. But it is of high uncertainty and needs to be set some parameters, such as Relief [12], genetic algorithm [37] and random subspace [22]. Heuristic search is designed to make a trade-off between high class-relevance and low features-redundancy. In which, greedy search strategy as a typical heuristic search performs a local search in feature space to find an optimum feature subset, such as forward greedy selection and backward elimination search. The forward greedy selection begins with a empty set, and adds one or several features with great significance into the set each time until the dependence does not increase. However, the backward search starts with the whole feature set and deletes one or several features with the significance as the dependence decreases. In real-word application, the heuristic search with reasonable computational costs is more feasible than the complete search and the stochastic search. Therefore, we employ the forward greedy sequential selection in this paper.

For the above mentioned feature selection algorithms, it is noticed that these algorithms focus on how to select high relevance and low redundancy features, but rarely consider the distinguishing ability of feature. In this paper, we propose a feature selection based on the quality of information (QIFS). First, we introduce classification margin to define the concept of maximum-nearest-neighbor, and integrate the concept of maximum-nearest-neighbor into Shannon's information theory. Then, we present maximum-nearest-neighbor entropy, maximum-nearest-neighbor joint entropy, and maximum-nearest-neighbor condition entropy, respectively. Correspondingly, we structure the class of maximum-nearest-neighbor, i.e., the distance of two samples less than or equal to the size of maximum-nearest-neighbor should be belonged to the same class; otherwise, the decision class is considered to be inconsistent. In addition, two samples belong to different classes, if there exists a feature can differentiate the maximum-nearest-neighbors of these two samples. Therefore, the concept of discrimination is proposed to distinguish samples, i.e., if the distance between two samples whose take the same class label is less than or equal to the maximum-nearest-neighbor of the sample, we can claim that the two samples are indistinguishable. Instead, if two samples have different labels and their distance is greater than the maximum-nearest-neighbor of the sample, we define that the two samples are distinguishable. Finally, a formula of evaluating the quality of feature is built.

In this paper, we present a feature selection method based on the maximum-nearest-neighbor information theory to estimate the quality of feature. It makes full use of the distinguishing ability of each feature, as well as fully considers the relevance and redundancy of each feature. To show the effectiveness of our feature selection method, we conduct a series of experiments. We first compare feature rank lists with respect to NRS [18]. Then, we show the classification accuracies of NRS, Relief [20], SPEC [46] and SPSF-LAR [48] with CART (Gini index), Linear Support Vector Machine (LSVM), and KNN, respectively. The main contributions of this paper are as follows:

- Identifying the quality of feature by distinguishing the between-class samples and the within-class samples.
- The relevant and non-redundant features can be successfully selected by the forward greedy sequential selection.
- The neighborhood parameter value is set by the margin of sample that avoiding the neighborhood parameter selection problem.
- Extensive experiments are conducted to show the effectiveness, stability, and scalability of the proposed method.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 proposes the related concepts of the Shannon's entropy and the neighborhood entropy. Section 4 presents the maximum-nearest-neighbor entropy for measuring feature's quality and constructs the corresponding algorithm. Experimental analysis is given in Section 5. Finally, conclusion is given in Section 6.

2. Related work

To date, a number of feature selection based on information theory methods have been developed [16,24,32,35,39,41]. In which, mutual information based feature selection is an effective evaluation criterion which possesses a solid theoretical foundation, i.e., mutual information can be used to analyze the upper and lower bound on the Bayes error rate [9]. Nevertheless, there are some long-standing challenges in statistics, such as the problems of evaluating high-dimensional joint mutual information and high-dimensional probability distribution. Therefore, some mutual information based feature selection methods having the difficulty that approximates the high-dimensional joint mutual information with low-dimensional mutual information. For the low-dimensional mutual information, 'relevancy' and 'redundancy' are popularly used to combination. For example, Battiti et al. [4] considered both the mutual information with respect to the output class and the already-selected features, and proposed the Mutual Information Feature Selection (MIFS). Peng et al. [32] used the mutual information as a metric to measure the relationship between the features and class via minimum-Redundancy-Maximum-Relevancy (mRMR) analysis. However, the two-dimensional mutual information can only find pairwise variable interactions, either between two features or between a feature and the class. Therefore, the two-dimensional mutual information cannot identify more complicated variable interactions. To address this problem, other techniques about the use of higher-dimensional mutual information quantities have been considered. Such as, Bennasar et al. [7] considered the interaction between the features and the classifier, and proposed two new non-linear feature selection methods, including Joint Mutual Information Maximisation (JMIM) and Normalised Joint Mutual Information Maximisation (NJMIM). Lin et al. [26] maximized the joint class-relevant information by explicitly reducing the class-relevant redundancies among features, and proposed a new algorithm called Conditional Informative Feature Extraction (CIFE). Nguyen et al. [30] provided a global solution for the mutual information based feature selection, and proposed Quadratic Programming Feature Selection (QPFS). Nguyen et al. [31] systematically investigated the issues of employing high-order dependencies for mutual information based features selection, and proposed a novel higher-order mutual information based feature selection(RelaxMRMR).

In addition, many extension of entropy have been presented to apply in different applications. Kosko [29] introduced a new non-probabilistic entropy measure in the context of fuzzy sets, and presented a framework of fuzzy information theory. Hu et al. [17] generalized Shannon's information entropy to neighborhood information entropy, and proposed the neighborhood information theory. Hu et al. [19] combined the advantage of robustness of Shannon's entropy with the ability of dominance rough sets in extracting ordinal structures, and proposed rank information theory. Zhang et al. [45] defined different fuzzy relations according to different types of features to measure the similarity between objects and in view of the effectiveness of entropy to measure information uncertainty, and proposed a fuzzy rough set-based information entropy theory.

3. Preliminary knowledge

3.1. Entropy and conditional entropy

In 1948, Shannon's entropy was first introduced as a measure of the uncertainty of random variables. Let U be the set of samples, $A \subseteq C$ is a subset of condition features and d is the decision feature. An equivalence relation R_A can be induced over U according to the values of samples on features A: $R_A = \{(x_i, x_j) | \forall\, a \in A, a(x_i) = a(x_j)\}$, where

$a(x)$ is the feature value of sample x on a. Then, the partition U/R_A generates a set of equivalence classes $\{X_1, X_2, \ldots, X_n\}$, where the elements in $X_{i(i=1,2,\ldots,n)}$ are indiscernible because they have the same feature values. Assume $X_1, X_2, \ldots, X_n$ are a set of random variables in U. The probability $p(X_i)$ of X_i is calculated as $|X_i|/|U|$, then Shannon's entropy of the partition is defined as

$$H(A) = -\sum_{i=1}^{n} p(X_i)\log p(X_i), \tag{1}$$

where $|X_i|$ denotes the cardinality of X_i.

It is noticed that Shannon's entropy does not rely on the real values, but the probabilities play a decisive role actually. Given another subset of features $B \subseteq C$, and B can induce the partition into $\{Y_1, Y_2, \ldots, Y_m\}$, then the joint entropy of features A and B is

$$H(A, B) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(X_i \cap Y_j)\log p(X_i \cap Y_j). \tag{2}$$

Assume that the subset of features B is given, called conditional entropy $H(A|B)$, which reflects the uncertainty of A, is defined as

$$\begin{aligned} H(A|B) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(X_i \cap Y_j)\log p(X_i|Y_j) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{|X_i \cap Y_j|}{|U|}\log\frac{|X_i \cap Y_j|}{|X_i|}. \end{aligned} \tag{3}$$

3.2. Neighborhood entropy and neighborhood conditional information

In order to compute information theory for hybrid data, we should know the probability distributions of variables and their joint distributions. However, these distributions are unknown in advance. Therefore, Hu et al. [17] presented an assumption that samples with the similar feature values should be classified into the same class or neighborhood class. Based on this assumption, the equivalent relation is extended into neighborhood relation. Moreover, Hu et al. [17] integrated the concept of neighborhood into Shannon's information theory, and proposed neighborhood information entropy, joint neighborhood entropy, and conditional neighborhood entropy, respectively.

Given a set of samples $U = \{x_1, x_2, \ldots, x_n\}$, we define Δ as a distance metric function on U, which satisfies $\Delta(x_i, x_j) \geq 0$; In general, Minkowski distance functions is defined as $\Delta_p(x_i, x_j) = [\sum_{l=1}^{n}(x_{li} - x_{lj})^p]^{1/p}$, such that $p = 1$, $p = 2$, and $p = \infty$, called Manhattan distance, Euclidean distance and Chebyshev distance, respectively. Given arbitrary $x_i \in U$ and feature space C, the neighborhood $\delta_C(x_i)$ of x_i in the feature space C is defined as

$$\delta_C(x_i) = \{x_j | x_j \in U, \Delta_C(x_i, x_j) \leq \delta\}, \tag{4}$$

where Δ is a metric function, $\forall\, x_1, x_2, x_3 \in U$, which satisfies

(1) $\Delta_C(x_1, x_2) \geq 0$, $\Delta_C(x_1, x_2) = 0$, if and only if $x_1 = x_2$;
(2) $\Delta_C(x_1, x_2) = \Delta_C(x_2, x_1)$;
(3) $\Delta_C(x_1, x_2) + \Delta_C(x_2, x_3) \geq \Delta_C(x_1, x_3)$.

Given two condition feature spaces A and B, $\delta_A(x)$ and $\delta_B(x)$ are the neighborhoods of x in A and B feature spaces, respectively. And it has: $\delta_{A\cup B}(x) = \delta_A(x) \cap \delta_B(x)$.

Given a set of samples $U = \{x_1, x_2, \ldots, x_n\}$, x_i described by hybrid feature set C, $A \subseteq C$ is a subset of feature set C. The neighborhood of sample x_i in A is defined as $\delta_A(x_i)$. Then the neighborhood uncertainty of the sample x_i is denoted by

$$NH_\delta^{x_i}(A) = -\log\frac{\| \delta_A(x_i)\|}{n}, \tag{5}$$

and the average uncertainty of all samples is computed as

$$NH_\delta(A) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\| \delta_A(x_i)\|}{n}. \tag{6}$$

Theorem 1 ([17]). If $\delta \leq \delta'$, $NH_\delta(A) \geq NH_{\delta'}(A)$.

Proof. $\forall\, x_i \in U$, we have $\delta(x_i) \subseteq \delta'(x_i)$, then $\| \delta(x_i)\| \leq \| \delta'(x_i)\|$, we have $NH_\delta(A) \geq NH_{\delta'}(A)$.□

Theorem 2 ([17]). If $\delta = 0$, *then* $NH_\delta(A) = H(A)$, where H(A) is Shannon's entropy.

Proof. If $\delta = 0$, the samples are classified into non-intersect $X_1, X_2, \ldots, X_m$, where $\Delta(x_i, x_j) = 0$ if $x_i, x_j \in X_k$. Assume there are m_i samples in X_i, then $H(A) = -\sum_{i=1}^{m}\frac{m_i}{n}log\frac{m_i}{n}$, and if $\delta = 0$, $x \in X_k$, then $\delta_A(x) = X_k$. If $i \neq j$, $X_i \cap X_j = \varnothing$, we have

$$\begin{aligned} NH_\delta(A) = -\frac{1}{n}log\frac{\| \delta_A(x_i)\|}{n} &= \sum_{x\in X_1} -\frac{1}{n}log\frac{\| \delta_A(x)\|}{n} + \cdots \\ &+ \sum_{x\in X_m} -\frac{1}{n}log\frac{\| \delta_A(x)\|}{n}. \end{aligned}$$

This is taken as proof that $NH_\delta(A) = H(A)$ if $\delta = 0$.□

Based on these Theorems 1 and 2, we know that neighborhood entropy is a natural generalization of Shannon's entropy. For numerical features, distance can be defined as $\Delta(x, y) = 0$ if x=y; otherwise, $\Delta(x, y) = 1$. If $\delta < 1$, the subset $\delta_A(x_i)$ of sample develops the equivalence class $[x_i]$, where $[x_i]$ is the set of sample having the same feature value with x_i, and the neighborhood entropy is equal to Shannon's entropy.

Assume $A, B \subseteq C$ are two subsets of features. The neighborhood of sample x_i in feature subspace $A \cup B$ is expressed as $\delta_{A\cup B}(x_i)$, then the joint neighborhood entropy [17] is

$$NH_\delta(A, B) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\| \delta_{A\cup B}(x_i)\|}{n}. \tag{7}$$

Particularly, if given C is a set of condition features and d is the class label. Then,

$$NH_\delta(C, d) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\| \delta_{C\cup d}(x_i)\|}{n}. \tag{8}$$

where $\delta_{C\cup d}(x_i) = \delta_C(x_i) \cap d_{x_i}$, and d_{x_i} is a set of samples which having the same class labels with x_i.

Assume $A, B \subseteq C$ are two subsets of features. The conditional neighborhood entropy [17] of A to B is defined as

$$NH_\delta(A|B) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\| \delta_{A\cup B}(x_i)\|}{\| \delta_B(x_i)\|}. \tag{9}$$

3.3. The measure of feature quality

Suppose that U_d is the subset of U consisting of samples with class d. Given an arbitrary sample $x \in U_d$, we call a sample $x' \in U - U_d$ is discriminated from x by a feature f if $x_f \neq x'_f$, where " − " represents the set subtraction [38]. Analogously, given a feature set C and a sample $x \in U_d$, for each sample $x' \in U - U_d$, there is a feature $f \in C$ such that $x_f \neq x'_f$, then we say that all samples in $U - U_d$ are discriminated from x by C.

Given $x \in U_d$, a feature f and a subset $U_d^{\sim} \subseteq U - U_d$, we say $Count(f, U_d^{\sim})$ to be the number of samples in $U_d^{\sim}$ that are discriminated from sample x by f [38]. In other words, the number of samples which have different classes from d and different values from x on f. Given a set of samples U, a feature set C, a labeled sample $x' \in U_{d(d=1,2,\ldots,T)}$, and a sample x. x' is claimed as the neighbor of x on C with class d if and only if for each feature $f \in C$, x and x' agree upon f: $x_f = x'_f$ [38].

For a feature f having p different values $\{1, 2, \ldots, p\}$, U_d^i shows the set of samples in U whose feature f takes i and class label is d, and U_*^i

expresses the union of $U_{d(i=1,2,\ldots,p;d=1,2,\ldots,T)}^{i}$ over d. The entropy of the feature f on U is defined as [33],

$$E(f, U) = -\sum_{i=1}^{p} \frac{|U_*^i|}{|U|} \sum_{d=1}^{T} \left(\frac{|U_d^i|}{|U_*^i|} \log_2 \left(\frac{|U_d^i|}{|U_*^i|} \right) \right). \tag{10}$$

Notice that the formula involves the logarithm of zero is defined as zero. The entropy of f offers a measure of impurity(uncertainty) for the subsets U_*^i of U ($i = 1, 2, \ldots, p$). Especially, if $E(f, U)$ is equal to zero, then U_*^i is pure(certainty) for $i = 1, 2, \ldots, p$, and a feature f can properly classify all samples in U.

Theorem 3. *Given a set of sample U, a condition feature f has p different values $\{1, 2, \ldots, p\}$, and d ($d = 1, 2, \ldots, T$) is the class feature, then $H(f|d) = E(f, U)$.*

Proof.

$$\begin{aligned} H(f|d) &= -\sum_{i=1}^{p} \sum_{d=1}^{T} p(X_i \cap U_d) \log_2 p(X_i|U_d) \\ &= -\sum_{i=1}^{p} \sum_{d=1}^{T} \frac{|X_i \cap U_d|}{|U|} \log_2 \frac{|X_i \cap U_d|}{|X_i|} \\ &= -\sum_{i=1}^{p} \frac{|X_i|}{|U|} \sum_{d=1}^{T} \frac{|X_i \cap U_d|}{|X_i|} \log_2 \frac{|X_i \cap U_d|}{|X_i|} \\ &= -\sum_{i=1}^{p} \frac{|U_*^i|}{|U|} \sum_{d=1}^{T} \left(\frac{|U_d^i|}{|U_*^i|} \log_2 \left(\frac{|U_d^i|}{|U_*^i|} \right) \right) = E(f, U). \end{aligned}$$

□

To make a trade-off between the impurity(uncertainty) of the partitions of U divided by a feature and the cardinality of a feature set, Wang et al. [38] presented the measure of feature quality. Now, when learning a feature set C for a sample $x \in U_d$, selecting the feature f whose $Q(f, U_d^{\sim}, U)$ is minimum among all unselected features in the current iteration. Given a sample $x \in U_d$, and a subset $U_d^{\sim} \subseteq U - U_d (d = 1, 2, \ldots, T)$, then the quality of f with regard to $U_d^{\sim}$ [38] is computed as

$$Q(f, U_d^{\sim}, U) = \begin{cases} +\infty & \text{if } Count(f, U_d^{\sim}) = 0 \\ \dfrac{E(f, U)}{Count(f, U_d^{\sim})} \times SI(f, U) & \text{otherwise,} \end{cases} \tag{11}$$

where $SI(f, U)$ is computed as follow,

$$SI(f, U) = -\sum_{i=1}^{p} \frac{|U_*^i|}{|U|} \log_2 \frac{|U_*^i|}{|U|}. \tag{12}$$

The term $SI(f, U)$ is presented by Quinlan [33] in C4.5, which is called the split information and is used to overcome the bias both the terms $E(f, U)$ and $Count(f, U_d^{\sim})$. It often biases the features whose have a large number of values.

Theorem 4. Given a set of samples U, a condition feature f has p different values $\{1, 2, \ldots, p\}$, then $H(f) = SI(f, U)$.

Proof.

$$\begin{aligned} H(f) &= -\sum_{i=1}^{p} p(X_i) \log_2 p(X_i) - \sum_{i=1}^{p} \frac{|X_i|}{|U|} \log_2 \frac{|X_i|}{|U|} - \sum_{i=1}^{p} \frac{|U_*^i|}{|U|} \log_2 \frac{|U_*^i|}{|U|} \\ &= SI(f, U). \end{aligned}$$

□

4. Feature selection based on quality of information

4.1. Margins and nearest neighbor

To process numeric data with information entropy directly, we introduce the concept of classification margin. There are many methods used to define the classification margin. From the viewpoint of the distance between sample and decision boundary [12], the classification margin can be defined as follow. Let P be the feature space, x be a sample, and Δ be a distance function. Then the large margin of x is defined as

$$\theta_P(x) = \Delta(x, NM(x)) - \Delta(x, NH(x)). \tag{13}$$

where $NM(x)$ is called the nearest miss, which denotes the nearest sample from different class of x, $NH(x)$ is called the nearest hit that denotes the nearest sample from the same class of x. $\Delta(x, NM(x))$ and $\Delta(x, NH(x))$ are the distance between x and its nearest sample from different and the same class, respectively.

Definition 1. Given arbitrary $x \in U$, C is hybrid feature space, $B \subseteq C$ is a subset of condition features. The maximum-nearest-neighbor $\eta(x)$ of x in feature space B is defined as

$$\eta(x) = \{x' | \Delta_B(x, x') \le d(x)), x' \in U\}. \tag{14}$$

where $d(x) = max(\Delta(x, NM(x)), \Delta(x, NH(x)))$.

Definition 2. Assume U_d is the subset of U consisting of samples with class d. Given $x \in U_d$, we define that the sample $x' \in U - U_d$ is discriminated from x by a feature f if $\Delta(x_f, x'_f) > d(x_f)$, where $d(x_f) = max(\Delta(x_f, NM(x_f)), \Delta(x_f, NH(x_f)))$, and " $-$ " represents the set subtraction. That is, x' is not the element of the maximum-nearest-neighbor $\eta(x)$ of x.

Similarly, given a feature set C and a sample $x \in U_d$, if for each sample $x' \in U - U_d$, there is a feature $f \in C$ such that $\Delta(x_f, x'_f) > d(x_f)$, then we say that all samples in $U - U_d$ are discriminated from x by C. In other words, if two samples belong to different classes, there is at least one feature, which leads to inconsistency with the maximum-nearest-neighbor of two samples.

Definition 3. Assume sample $x \in U_d$, a condition feature f and a subset $U_d^{\sim} \subseteq U - U_d$, $x' \in U_d^{\sim}$, we indicate $Count_\eta(f, U_d^{\sim})$ as the number of samples in $U_d^{\sim}$ that are discriminated from x by f. It is equal to the cardinality of $\Delta(x_f, x'_f) > d(x_f)$.

Definition 4. For a set of samples U, a condition feature set C, a labeled sample $x' \in U_{d(d=1,2,\ldots,T)}$, and a sample x, we say that x' is the maximum-nearest-neighbor of x on C with class d if and only if for each feature $f \in C$, x and x' take agree upon f: $x'_f \in \eta(x_f)$.

4.2. Maximum-nearest-neighbor entropy and quality of information

For dealing with the hybrid data directly, we present the concept of the maximum-nearest-neighbor by employing large margin. Then, we generalize Shannon's information entropy to maximum-nearest-neighbor information entropy, and present maximum-nearest-neighbor entropy and maximum-nearest-neighbor conditional entropy, respectively.

Definition 5. Given a set of samples $U = \{x_1, x_2, \ldots, x_n\}$ described by hybrid features C, $A \subseteq C$ is a subset of features. The maximum-nearest-neighbor of sample x_i in A is computed as $\eta_A(x_i)$. Then the maximum-nearest-neighbor uncertainty of the sample x_i is defined as

$$MH_\eta^{x_i}(A) = -\log \frac{\| \eta_A(x_i) \|}{n}, \tag{15}$$

and the average uncertainty of all samples is

Table 1
Dataset description.

Dataset	Abbreviation	Samples	Features	Labels
Australian Credit	Australian	690	14	2
Credit Approval	Credit	690	15	2
Dermatology	Derm	366	34	6
Glass Identification	Glass	240	9	7
Horse Colic	Horse	368	22	2
Ionosphere	Iono	351	34	2

Table 2
Subsets of feature selected in different algorithms.

Dataset	N1	QIFS	N2	NRS
Australian	8✓	8,12,11,4,5,9,1,14	12	12,13,7,11,6,1,4,3,2,8,5,10
Credit	9✓	9,10,12,1,6,13,5,4,7	14	11,13,4,12,7,1,14,10,3,2,9,6,8,15
Derm	11	11,31,22,12,9,15,25,5,8,14,7	10✓	18,1,32,16,31,4,15,5,27,22
Glass	5✓	4,1,3,9,8	8	2,1,3,9,5,7,4,8
Horse	4✓	15,2,1,20	7	4,19,12,10,20,17,5
Iono	4✓	1,5,3,6	7	3,9,24,34,13,5,1

Table 3
Classification accuracies (%) of different feature selection algorithms with CART.

Dataset	Raw	NRS	Relief	SPSF-LAR	SPEC	QIFS
Australian	82.60 ± 4.45	**83.63 ± 4.09**	82.31 ± 4.39	82.63 ± 4.39	81.31 ± 4.77	82.90 ± 5.66
Credit	82.73 ± 14.86	82.30 ± 14.74	82.59 ± 14.74	82.44 ± 14.84	82.29 ± 14.70	**83.47 ± 16.31**
Derm	92.26 ± 6.69	93.49 ± 4.97	77.18 ± 7.67	81.15 ± 10.29	85.91 ± 6.33	**94.21 ± 5.17**
Glass	43.62 ± 15.68	44.01 ± 18.55	46.84 ± 17.09	49.70 ± 17.23	45.53 ± 16.40	**50.63 ± 19.99**
Horse	95.92 ± 2.30	90.76 ± 4.87	94.81 ± 3.06	85.61 ± 5.08	73.67 ± 5.85	**95.91 ± 2.71**
Iono	87.55 ± 6.93	89.46 ± 4.41	84.42 ± 8.30	84.97 ± 7.60	83.90 ± 7.49	**90.98 ± 6.76**
Average	80.78	80.61	78.02	77.75	75.44	**83.02**

Table 4
Classification accuracies (%) of different feature selection algorithms with LSVM.

Dataset	Raw	NRS	Relief	SPSF-LAR	SPEC	QIFS
Australian	85.52 ± 5.20	**85.52 ± 5.20**	**85.52 ± 5.20**	**85.52 ± 5.20**	**85.52 ± 5.20**	**85.52 ± 5.20**
Credit	85.48 ± 18.51	**85.48 ± 18.51**	**85.48 ± 18.51**	**85.48 ± 18.51**	73.35 ± 6.00	**85.48 ± 18.51**
Derm	96.55 ± 2.82	93.17 ± 5.42	77.68 ± 5.19	85.87 ± 4.57	84.84 ± 5.77	**95.32 ± 4.37**
Glass	57.11 ± 11.57	58.42 ± 10.07	56.22 ± 12.08	**58.47 ± 11.09**	57.11 ± 10.96	58.02 ± 11.03
Horse	92.96 ± 4.43	88.29 ± 4.50	87.22 ± 4.96	89.38 ± 5.20	76.68 ± 6.48	**89.39 ± 3.95**
Iono	87.57 ± 6.45	86.68 ± 5.89	**86.69 ± 5.87**	74.99 ± 8.66	66.12 ± 3.27	**86.69 ± 5.87**
Average	84.20	82.93	80.14	79.95	75.96	**83.40**

Table 5
Classification accuracies (%) of different feature selection algorithms with KNN(K=5).

Dataset	Raw	NRS	Relief	SPSF-LAR	SPEC	QIFS
Australian	85.37 ± 3.81	84.78 ± 3.22	84.93 ± 4.21	84.20 ± 4.48	**85.36 ± 4.50**	85.09 ± 4.88
Credit	82.89 ± 16.12	82.59 ± 16.63	**83.76 ± 15.73**	**83.76 ± 15.73**	83.03 ± 16.23	83.47 ± 16.92
Derm	97.10 ± 2.91	90.75 ± 7.55	80.79 ± 1.94	84.21 ± 5.05	81.83 ± 7.73	**95.91 ± 2.67**
Glass	65.49 ± 9.13	65.49 ± 9.13	68.64 ± 16.59	64.58 ± 10.50	65.97 ± 7.24	**68.68 ± 8.29**
Horse	90.24 ± 4.76	89.11 ± 43.45	91.57 ± 3.25	88.59 ± 5.22	73.17 ± 10.74	**93.23 ± 5.50**
Iono	64.12 ± 1.09	91.80 ± 5.39	64.12 ± 1.09	82.92 ± 7.02	64.12 ± 1.09	**92.38 ± 4.28**
Average	80.87	84.09	78.97	81.37	75.958	**86.46**

$$MH_\eta(A) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\eta_A(x_i)\|}{n}. \tag{16}$$

Definition 6. Assume $A, B \subseteq C$ are two subsets of features. The maximum-nearest-neighbor of sample x_i in feature subspace $A \cup B$ is defined as $\eta_{A\cup B}(x_i)$, then we define the joint maximum-nearest-neighbor entropy as

$$MH_\eta(A, B) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\eta_{A\cup B}(x_i)\|}{n}. \tag{17}$$

Definition 7. Assume $A, B \subseteq C$ are two subsets of features. The maximum-nearest-neighbor conditional entropy of A to B is defined as

$$MH_\eta(A|B) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\eta_{A\cup B}(x_i)\|}{\|\eta_B(x_i)\|}. \tag{18}$$

where if C is a set of condition features and d is the class feature, and $\eta_{A\cup C}(x_i) = \eta_A(x_i) \cap d_{x_i}$, we have

$$MH_\eta(C|d) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\eta_{C\cup d}(x_i)\|}{\|\eta_d(x_i)\|}. \tag{19}$$

According to Theorems 3 and 4, the maximum-nearest-neighbor entropy and maximum-nearest-neighbor conditional entropy can be used to propose a new measure for evaluating the quality of feature in Definition 8.

Definition 8. Given a sample $x \in U_d$, and a subset of samples $U_{\tilde{d}} \subseteq U - U_d (d = 1, 2, \ldots, T)$, then the quality of feature f with regard to $U_{\tilde{d}}$ is defined as,

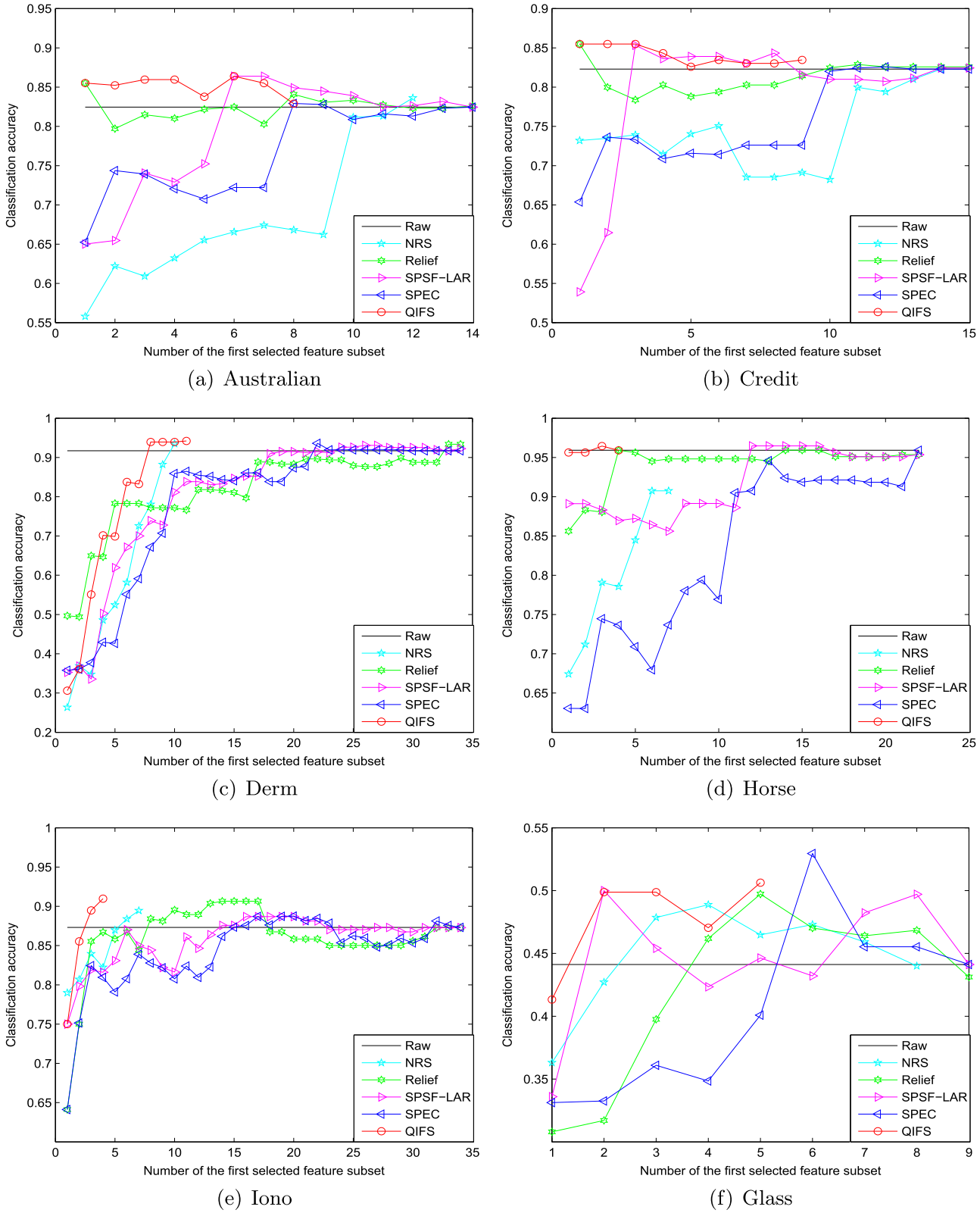


Fig. 1. Classification accuracy vs. number of the selected feature subset (CART). (a) Australian (b) Credit (c) Derm (d) Horse (e) Iono (f) Glass.

$$Q_\eta(f, U_d^{\sim}, U) = \begin{cases} +\infty & \text{if } Count_\eta(f, U_d^{\sim}) = 0 \\ \dfrac{MH_\eta(f|d)}{Count_\eta(f, U_d^{\sim})} \times MH_\eta(f) & \text{otherwise} \end{cases} \tag{20}$$

4.3. Algorithm

The objective of feature selection is to find a subset of features which has the same discriminating power as the original data. With the proposed measure, a forward greedy search algorithm for feature selection can be formulated. The pseudo-code is described in

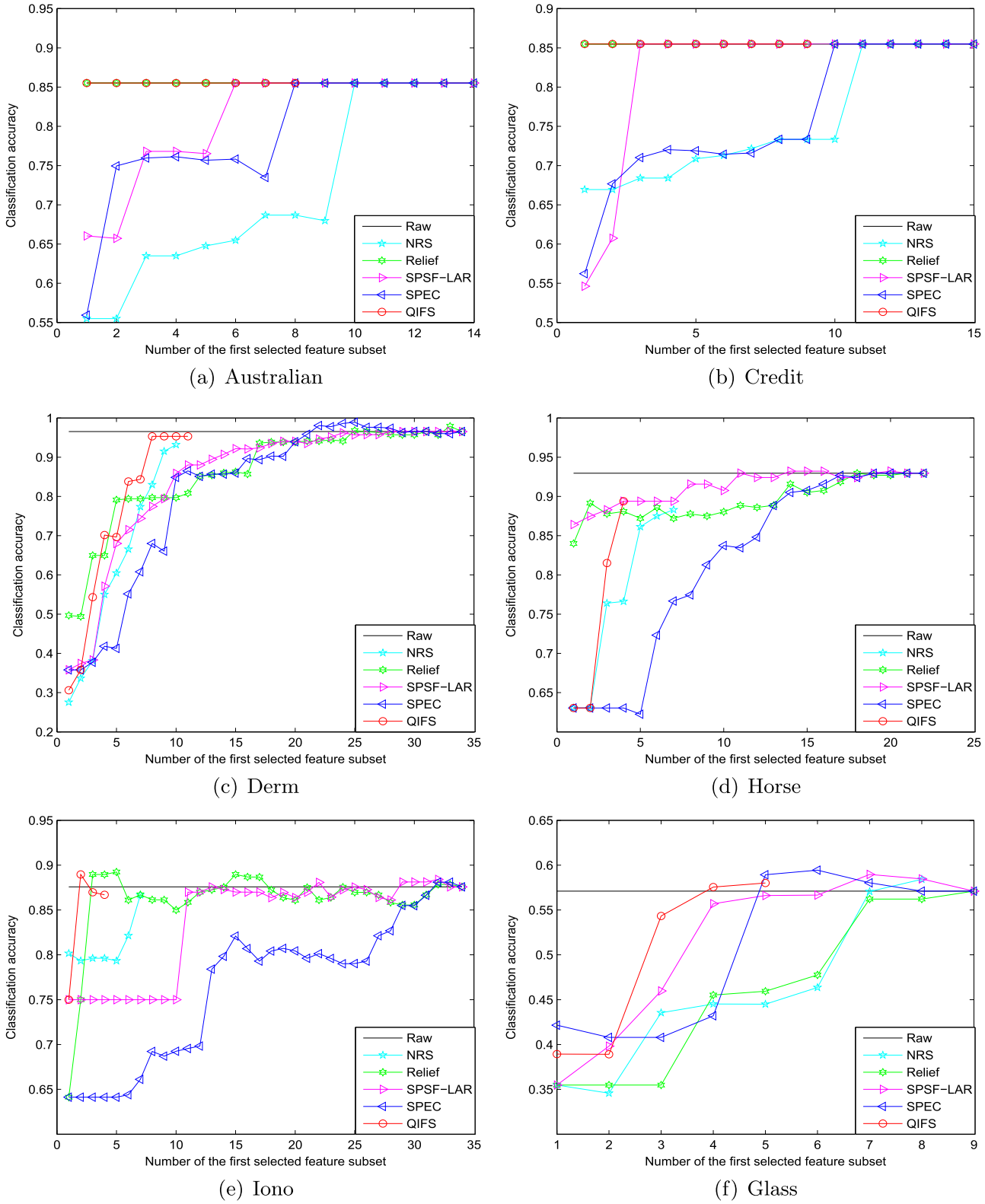


Fig. 2. Classification accuracy vs. number of the selected feature subset (LSVM). (a) Australian (b) Credit (c) Derm (d) Horse (e) Iono (f) Glass.

Algorithms 1 and 2, respectively.

Algorithm 1. Intra_class_Feature_Select.

Input: U: a set of samples; C: a set of condition features; d: the class feature; n: the cardinality of C; T: the cardinality of d.

Output: *FSet*: the subset of final selected features.

1: **for** C=1 to n **do**

2: for each sample x, compute the maximum-nearest-neighbor $\eta(x)$ and the distance $D(x)$ of x, the maximum-nearest-neighbor entropy $MH_\eta(C)$, the maximum-nearest-neighbor conditional entropy $MH_\eta(C|d)$, the product with $MH_\eta(C)$ and $MH_\eta(C|d)$,

Table 6
Classification accuracies (%) comparison on noisy datasets with CART.

Dataset	NRS	Relief	SPSF-LAR	SPEC	QIFS
Australian	75.56	79.88	73.82	66.69	**82.21**
Credit	82.15	80.24	**83.43**	70.94	82.22
Derm	64.34	64.11	37.03	39.73	**68.42**
Glass	64.19	75.75	64.08	**80.82**	80.13
Horse	89.07	90.46	88.07	81.99	**92.65**
Iono	84.33	84.48	84.91	79.76	**87.27**
Average	76.60	79.15	71.89	69.99	**82.15**

Table 7
Classification accuracies (%) comparison on noisy datasets with LSVM.

Dataset	NRS	Relief	SPSF-LAR	SPEC	QIFS
Australian	77.46	80.94	76.81	63.29	**83.55**
Credit	85.48	85.48	85.48	70.94	**85.48**
Derm	64.51	65.06	40.48	39.53	**68.27**
Glass	53.25	51.76	55.89	51.36	**57.92**
Horse	86.96	85.82	86.44	63.04	**87.66**
Iono	**84.64**	80.97	81.18	64.72	82.03
Average	75.39	75.01	71.05	58.75	**77.49**

Table 8
Classification accuracies (%) comparison on noisy datasets with KNN.

Dataset	NRS	Relief	SPSF-LAR	SPEC	QIFS
Australian	71.31	65.65	71.63	52.26	**75.47**
Credit	82.85	82.20	**83.56**	65.97	82.36
Derm	63.66	64.18	36.91	39.63	**68.57**
Glass	47.38	46.19	47.73	44.26	**50.20**
Horse	89.21	90.71	86.80	76.40	**92.38**
Iono	76.73	86.11	86.00	65.51	**88.52**
Average	71.86	72.51	68.77	57.34	**76.25**

expressing as MMH;
end for
initialize $FSet = \varnothing$;
for d= 1 to T **do**
  $Temp \longleftarrow U_d$;
  while ($Temp \neq \varnothing$) **do**
    $U_d^{\sim} \longleftarrow U - U_d$;
    $x \in Temp$;
    for C= 1 to n **do**
      compute $Count_\eta(C, U_d^{\sim})$;
    end for
    Inter_FS = Inter_class_Feature_Select $(x, MMH, Count_\eta(C, U_d^{\sim}), U_d^{\sim}, D(x))$;
    if ($InterFS \cap FSet \neq InterFS$)
      $FSet \longleftarrow FSet \cup InterFS$;
    end if
    for $FSet$ = 1 to $|FSet|$ **do**
      delete the maximum-nearest-neighbors of x from $Temp$;
    end for
  end while
end for

Algorithm 2. Inter_class_Feature_Select.

Input: $(x, MMH, Count_\eta(C, U_d^{\sim}), U_d^{\sim}, D(x))$.
Output: $InterFS$: the subset of Inter_class_Feature_Select.

initialize $InterFS \longleftarrow \varnothing$;
$Q_\eta = \varnothing$;
$FeatureSet = \{C_1, C_2, \ldots, C_{column-1}\}$;
for C= 1 to n **do**
  compute $Q_\eta(C, U_d^{\sim}, U)$ usingDefinition 8;
end for
while ($U_d \neq \varnothing$ and $FeatureSet \neq \varnothing$) **do**
  $Count_{min} \longleftarrow argmin_{C \in FeatureSet} Q_\eta(C, U_d^{\sim}, U)$;
  $InterFS \longleftarrow InterFS \cup \{Count_{min}\}$;
  $FeatureSet \longleftarrow FeatureSet - \{Count_{min}\}$;
  delete the samples discriminated of x from $U_d^{\sim}$;
  for C= 1 to $FeatureSet$ **do**
    compute $Count_\eta(C, U_d^{\sim})$;
    compute $Q_\eta(C, U_d^{\sim}, U)$ usingDefinition 8;
  end for
end while

Algorithm 1 is a forward feature selection process, i.e., it begins with an empty set *FSet* of feature, and adds a minimum feature subset which can effectively distinguish and identify a given sample x with its between-class samples and the within-class samples into the set *FSet* in each round. In order to avoid selecting redundant features, we delete the repeated features in each round. This is a strategy of subset generation. Then, we embed the subset evaluation in this strategy by minimizing Eq. (20). Algorithm 1 does not stop until all samples are distinguished. According to the feature quality function and the strategy of search, we can understand that the algorithm tries to find a feature subspace such that there is the least overlapped region between classes for a given classification task.

5. Experiments

In this section, we carry out several experiments to study the effectiveness of our approach. We first give a brief description of datasets and experimental settings. Then, we empirically compare the proposed method with other state-of-the-art feature selection algorithms.

5.1. Experimental datasets and settings

To validate the proposed algorithm, we select six datasets from UCI Repository of machine learning databases [3]. Table 1 displays certain standard statistics of datasets such as the number of samples, features, and labels, respectively.

In our experiments, we compare the proposed method with other four popular feature selection algorithms: NRS [18], Relief [20], SPEC [46] and SPSF-LAR [48]. To verify the efficacy of different algorithms more specifically and clearly, we select the same number of features with the quantity determined by NRS as the final feature subset, due to NRS gets a feature subset directly. In addition, the neighborhood size δ of NRS is set as 0.1, as recommended in [18]. To validate and evaluate the classification performance with impartial results, we adopt 10-fold-cross validation for each algorithm-dataset combination when testing the ability of classification. Finally, three base classifiers: CART, LSVM, and KNN are introduced to test the quality of the selected features, respectively. In which, we select the Gini index as a split criterion for CART, use the linear kernel for SVM, and set K=5 for KNN, respectively.

5.2. Classification performance comparison

To show the compactness of our proposed algorithm, we take NRS as the compared algorithm due to NRS gets feature subset directly. The selected feature subset is shown in Table 2, in which the second and

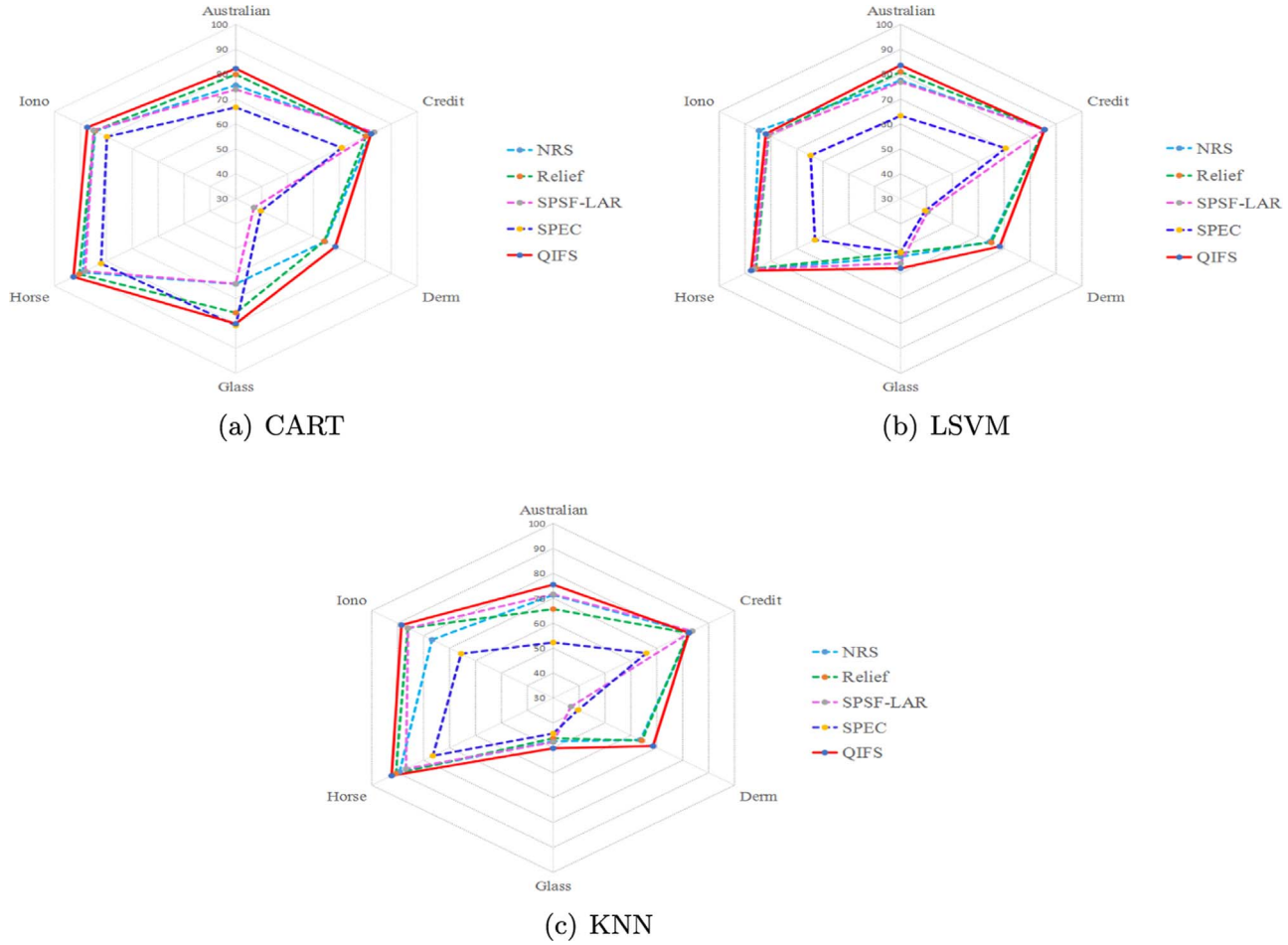


Fig. 3. Spider web diagram showing the classification accuracy to compare the stability index on noisy datasets with CART, LSVM, and KNN. (a) CART (b) LSVM (c) KNN.

Table 9
Dataset description.

Dataset	Samples	Features	Labels
Waveform	5000	21	3
Sick	2800	29	2
PIE10P	210	2420	10
YALE	165	1024	15

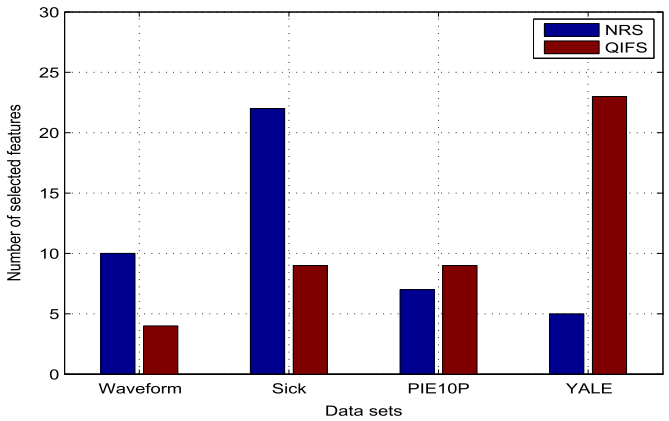


Fig. 4. Number of selected features on different datasets.

fourth columns represent the number of selected features with QIFS and NRS, respectively, and the third column is the subset of selected features obtained by QIFS. Accordingly, the fifth column denotes the sequence of feature subset with NRS. Symbol "✓" in Table 2 means the number of selected features is minimal.

From Table 2, we observe that (1) The selected feature lists from NRS and QIFS are different; (2) QIFS obtains minimal feature subset on five datasets, the main reason for this results is that QIFS can find a feature subspace having the least overlapped region between different classes, and there is a little any redundant between features.

To show the effectiveness of feature selection, we compare the classification accuracies among NRS, Relief, SPEC, and SPSF-LAR. The classification performance of different algorithms are shown in Tables 3–5, where the Raw denotes the classification accuracy on original dataset, the last line in italics represents the average classification accuracy, and bold font means the best classification accuracy for each dataset. From the results in Tables 3–5, it can be observed that QIFS gets the best classification accuracy on five datasets with CART and LSVM, and four datasets with KNN, respectively. Especially, as to CART, QIFS significantly outperforms the original classification accuracy on all datasets; as to LSVM and KNN, QIFS achieves superior comparable performance against the original data on four datasets and gets at least comparable performance on other two datasets. In addition, Tables 3–5 show that QIFS dramatically improves the classification performance compared with other four feature selection algorithms, and achieves the highest average classification accuracy with CART, LSVM and KNN, respectively.

In order to fully explain the change situation of performance between QIFS and the comparing algorithms, we only select two base classifiers to show the classification performance with different datasets after feature selection, according to the paper's compactness. Figs. 1 and 2 plot the classification accuracy curves of different algorithms over

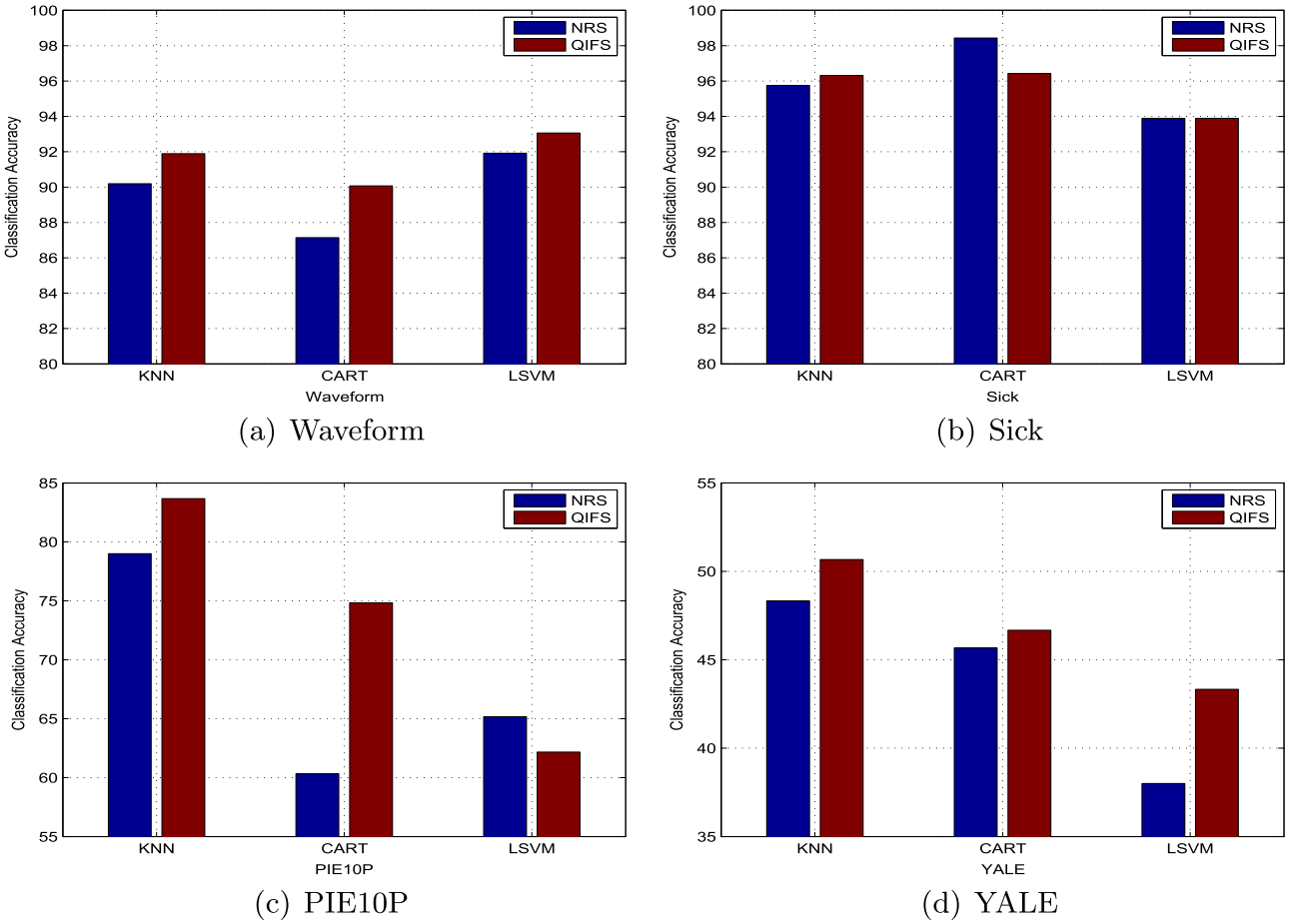


Fig. 5. Classification performance comparison on different datasets with different classifiers. (a) Waveform (b) Sick (c) PIE10P (d) YALE.

six datasets based on different classifiers. In these figures, the horizontal and vertical axes represent the number of selected features and the classification accuracy, respectively. The results in these figures demonstrate the situation in classification accuracy with the number of selected features. Note that the black line denotes the final classification accuracy of Raw data, which provides a base line to evaluate the performance of different feature selection algorithms. For making figures clearly, the standard deviation is not given in these figures. We can observe that QIFS is almost consistently better than other four feature selection algorithms on all six datasets with CART classifier, and on four datasets with LSVM classifier, respectively. Interestingly, we can find that the classification accuracies do not monotonously increase with the number of selected features. Instead, with the increase of features, the accuracies increases firstly, and then stabilize or reduce gradually after reaching a peak value. These results show that it is non-trivial to conduct feature selection to obtain a compact subset of superior features. From Figs. 1 and 2, we can conclude that QIFS is most better than NRS, Relief, SPEC, and SPSF-LAR with different classifiers.

5.3. Stability analysis

To validate the stability of the proposed algorithm, we add noise to the feature values. First, we generate $n \times m$ (where n denotes the number of samples and m denotes the number of features) noisy data that satisfy a Gaussian distribution, and multiply them by θ (where the value of θ is selected randomly in this paper). In which, we carry out experiment for 10 times with each dataset, and then use the average value of classification accuracy as the final result after ten random runnings. For the feature with noisy data, the classification accuracy with different classifiers after feature selection are shown in Tables 6–8. Tables 6–8 show that the average classification accuracy of QIFS outperforms the other comparing algorithms. In addition, QIFS achieves statistically superior comparable performance against other comparing algorithms on five out of six datasets, and obtains at least comparable performance on the other one dataset with CART and KNN.

To verify the stability of different algorithms more specifically and clearly, we draw spider web diagram to show the stability index on noisy datasets. Fig. 3 shows the classification accuracy to compare the stability index on noisy datasets with CART, LSVM, and KNN, respectively. These results show that QIFS can obtain optimum values more stable than other comparing algorithms on all datasets.

5.4. Scalability analysis

To demonstrate the scalability of our method, we conduct another series of experiments. In this part, we use four datasets from different domains, as outlined in Table 9. In which, Waveform and Sick include less features but adequate samples, but PIE10P and YALE are high-dimensional and small-sample datasets. These four datasets reflect two characteristics from the ratio between samples and features.

For showing the comparative results impartially, we compare the effectiveness of QIFS with NRS due to these two algorithms obtain feature subset directly, and we divide the experiments into two parts. In the first part, we compare the number of selected features on different datasets, as shown in Fig. 4. In the second part, we compare the classification performance on different datasets with different classifiers, as shown in Fig. 5. From Figs. 4 and 5, we can observe that: (1) from Fig. 4, QIFS selects a fewer features than NRS on

Waveform and Sick, but more than NRS on PIE10P and YALE. For Waveform and Sick, the number of selected features meet the compactness of feature selection. The reason for the results on PIE10P and YALE are that these two datasets with high-dimensional feature space. In addition, if the number of selected features on PIE10P and YALE are too small, it will miss some valuable features and is difficult to recognize samples under the small feature space. (2) From Fig. 5, the classification accuracy of QIFS outperforms NRS on all datasets with different classifiers, except for Sick with CART and PIE10P with LSVM. To summarize, QIFS achieves highly competitive performance against NRS, the number of selected features is more reasonable than NRS, and QIFS has scalability better than NRS.

6. Conclusions

In this paper, we presented a new feature selection method based on the quality of information, which pays more attention to the distinguishing ability of feature itself. We first introduced classification margin to form the concept of maximum-nearest-neighbor, employed the maximum-nearest neighbor to generalize Shannon's information theory, and proposed the maximum-nearest-neighbor information theory for processing hybrid data. Then, we defined the criterion of discrimination with sample by employing the maximum-nearest-neighbor. Finally, we constructed a formula for evaluating the quality of feature. Experimental results show that the proposed algorithm is more effective than many popular feature selection methods. In addition, some other experiments also demonstrate the scalability and stability of the proposed method.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (Nos. 61303131 and 61672272), the Natural Science Foundation of Fujian Province (No. 2013J01028), the Department of Education, Fujian Province under grant (No. JA14192), the China Postdoctoral Science Foundation (2015M581298), and the Program for New Century Excellent Talents in Fujian Province University.

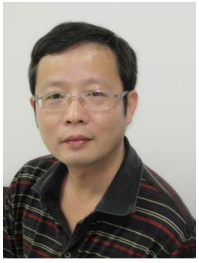
References

[1] A.-A. Ahmed, A dependency-based search strategy for feature selection, Expert Syst. Appl. 10 (36) (2009) 12392–12398.
[2] H. Almuallim, T. Dietterich, Learning with many irrelevant features, in: Proceedings of the 9th National Conference on Artificial Intelligence. Menlo Park, Vol. 2, 1991, pp: 547–552.
[3] K. Bache, M. Lichman, UCI Machine Learning Repository, University of California, School of Information and Computer Science: Irvine, CA, 2013. (http://archive.ics.uci.edu/ml).
[4] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Netw. 5 (4) (1994) 537–550.
[5] R. Bellman, Adaptive Control Processes: A Guided Tour, Princeton University Press, Princeton, New Jersey, USA, 1961.
[6] P. Bermejo, J. Gamez, J. Puerta, Speeding up incremental wrapper feature subset selection with Naive Bayes classifier, Knowl.-Based Syst. 55 (2014) 140–147.
[7] M. Bennasar, H. Yulia, S. Rossitza, Feature selection using Joint Mutual Information Maximisation, Expert Syst. Appl. 42 (22) (2015) 8520–8532.
[8] L. Breiman, J. Friedman, C. Stone, et al., Classification and Regression Trees, CRC Press: Boca Raton, 1984.
[9] G. Brown, A. Pocock, M. Zhao, et al., Conditional likelihood maximisation: a unifying framework for information theoretic feature selection, J. Mach. Learn. Res. 13 (1) (2012) 27–66.
[10] G. Chen, J. Chen, A novel wrapper method for feature selection and its applications, Neurocomputing 159 (2015) 219–226.
[11] L. Chiang, R. Pell, Genetic algorithms combined with discriminant analysis for key variable identification, J. Process Control 14 (2) (2004) 143–155.
[12] R. Gilad–Bachrach, A. Navot, N. Tishby, Margin based feature selection–theory and algorithms, in: Proceedings of the 21st International Conference on Machine Learning, Vol. 6184, 2004, pp: 40–48.
[13] I. Guyon, A. Elisseeff, An introduction to variable and features election, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[14] I. Guyon, J. Weston, S. Barnhill, et al., Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1/2/3) (2002) 389–422.
[15] M. Hall, Correlation–based feature selection for discrete and numeric class machine learning, in: Proceedings of the 17th international Conference on machine learning, May 2000, pp: 359–366.
[16] N. Hoque, D. Bhattacharyya, J. Kalita, MIFS-ND: a mutual information-based feature selection method, Expert Syst. Appl. 41 (2014) 6371–6385.
[17] Q. Hu, L. Zhang, D. Zhang, et al., Measuring relevance between discrete and continuous features based on neighborhood mutual information, Expert Syst. Appl. 38 (2011) 10737–10750.
[18] Q. Hu, D. Yu, J. Liu, et al., Neighborhood rough set based heterogeneous feature subset selection, Inf. Sci. 178 (18) (2008) 3577–3594.
[19] Q. Hu, X. Che, L. Zhang, et al., Rank entropy-based decision trees for monotonic classification, IEEE Trans. Knowl. Data Eng. 24 (11) (2012) 2052–2064.
[20] K. Kira, L. Rendell, The feature selection problem: Traditional methods and a new algorithm, in: Proceedings of the 9th National Conference on Artificial Intelligence, Menlo Park, 1992, pp: 129–134.
[21] I. Kononenko, Estimation attributes: Analysis andextensions of RELIEF, in: Proceedings of the 1994 European Conference on Machine Learning, New Brunswick, 1994, pp: 171–182.
[22] C. Lai, M. Reinders, L. Wessels, Random subspace method for multivariate feature selection, Pattern Recognit. Lett. 27 (2006) 1067–1076.
[23] Y. Lin, J. Li, P. Lin, Feature selection via neighborhood multi-granulation fusion, Knowl.-Based Syst. 67 (2014) 162–168.
[24] Y. Lin, X. Hu, X. Wu, Quality of information-based source assessment and selection, Neurocomputing 133 (2014) 95–102.
[25] Y. Lin, Q. Hu, J. Liu, et al., Multi-label feature selection based on max-dependency and min-redundancy, Neurocomputing 168 (2015) 92–103.
[26] D. Lin, X. Tang, Conditional infomax learning: an integrated framework for feature extraction and fusion, Computer Vision CECCV 2006. Springer: Berlin, Heidelberg, 2006, pp. 68–82.
[27] H. Liu, X. Wu, S. Zhang, Feature selection using hierarchical feature clustering, in: Proceedings of the 20th ACM International Conferenceon Information and Knowledge Management, 2011, pp: 979–984.
[28] H. Liu, L. Yu, Toward integrating features election algorithms for classification and clustering, IEEE Trans. Knowl. Data Eng. 17 (4) (2005) 491–502.
[29] B. Kosko, Fuzzy entropy and conditioning, Inf. Sci. 40 (2) (1986) 165–174.
[30] X. Nguyen, J. Chan, S. Romano, et al., Effective global approaches for mutual information based feature selection, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, NewYork, NY, USA, 2014, pp: 512C521.
[31] X. Nguyen, S. Zhou, J. Chan, et al., Can high-order dependencies improve mutual information based feature selection?, Pattern Recognit. 53 (2016) 46–58.
[32] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[33] J. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufman: San Mateo, 1993.
[34] J. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.
[35] W. Qian, W. Shu, Mutual information criterion for feature selection from incomplete data, Neurocomputing 168 (2015) 210–220.
[36] Z. Sun, B. George, M. Ronald, Object detection using feature subset selection, Pattern Recognit. 37 (11) (2004) 2165–2176.
[37] H. Vafaie, I. Imam, Feature selection methods: genetic algorithms vs. greedy-like search, in: Proceedings of International Conference on Fuzzy and Intelligent Control Systems, 1994.
[38] J. Wang, P. Zhang, G. Wen, Classifying categorical data by rule–based neighbors, in: Proceedings of the 11th IEEE International Conference on Data Mining, 2011, pp: 1248–1253.
[39] F. Wang, J. Liang, An efficient feature selection algorithm for hybrid data, Neurocomputing 193 (2016) 33–41.
[40] Z. Wang, B. Zineddin, J. Liang, et al., cDNA microarray adaptive segmentation, Neurocomputing 142 (2014) 408–418.
[41] M. Wei, T. Chow, R. Chan, Heterogeneous feature subset selection using mutual information-based feature transformation, Neurocomputing 168 (2015) 706–718.
[42] L. Yu, H. Liu, Feature selection for high–dimensional data: a fast correlation–based filter solution, in: Proceedings of International Conference on Machine Learning (ICML'03), 2003, pp: 856–863.
[43] L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learn. Res. 5 (1) (2004) 1205–1224.
[44] F. Yue, W. Zuo, Consistency analysis on orientation features for fast and accurate palmprint identification, Inf. Sci. 268 (2014) 78–90.
[45] X. Zhang, C. Mei, D. Chen, et al., Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy, Pattern Recognit. 56 (2016) 1–15.
[46] Z. Zhao, H. Liu, Spectral Feature Selection for Supervised and Unsupervised Learning, in: Proceedings of the 24th International Conference on Machine Learning (ICML), 2007, pp: 1151–1157.
[47] H. Zhao, K. Qin, Mixed feature selection in incomplete decision table, Knowl.-Based Syst. 57 (2014) 181–190.
[48] Z. Zhao, L. Wang, H. Liu, et al., On similarity preserving feature selection, IEEE Trans. Knowl. Data Eng. 24 (3) (2013) 619–632.
[49] Z. Zhu, Y. Ong, M. Dash, Wrapper-filter feature selection algorithm using a memetic framework, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 37 (2007) 70–76.
[50] W. Zhu, G. Si, Y. Zhang, et al., Neighborhood effective information ratio for hybrid feature subset evaluation and selection, Neurocomputing 99 (2013) 25–37.

Jinghua Liu received the M.S. degree from the School of Computer Science, Minnan Normal University. She currently is a Ph.D. student in Department of Automation, Xiamen University. Her research interests are focused on data mining and granular computing.

Yaojin Lin received the Ph.D. degree in School of Computer and Information from Hefei University of Technology. He currently is an associate professor with Minnan Normal University and a postdoctoral fellow with Tianjin University. His research interests include data mining, and granular computing. He has published more than 50 papers in many journals, such as Neurocomputing, Decision Support Systems, Information Sciences, and Applied Intelligence.

Menglei Lin received the M.S. degree in Mathematics from Xiamen University. He currently is a professor in School of Computer Science, Minnan Normal University. His research interests include granular computing.



Shunxiang Wu received the M.S. degree in Department of Computer Science and Engineering from Xi'an Jiaotong University in 1991 and the Ph.D. degree in School of Economics and Management, Nanjing University of Aeronautics & Astronautics in 2007. He is currently a professor in Department of Automation, Xiamen University. His research interests include intelligent computing, data mining, and granular computing.

Jia Zhang is currently working toward the Master degree from the School of Computer Science, Minnan Normal University. He research interests are focused on data mining.