Pattern Recognition 45 (2012) 2992–3002

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Feature evaluation and selection with cooperative game theory

Xin Sun [a,b], Yanheng Liu [a,b,*], Jin Li [c], Jianqi Zhu [a,b], Huiling Chen [a,b], Xuejie Liu [a,b]

[a] College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China
[b] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China
[c] School of Philosophy and Society, Jilin University, Changchun, Jilin 130012, China

ARTICLE INFO

Article history:
Received 3 June 2011
Received in revised form
12 January 2012
Accepted 2 February 2012
Available online 10 February 2012

Keywords:
Machine learning
Feature selection
Cooperative game theory
Filter method

ABSTRACT

Recent years, various information theoretic based measurements have been proposed to remove redundant features from high-dimensional data set as many as possible. However, most traditional Information-theoretic based selectors will ignore some features which have strong discriminatory power as a group but are weak as individuals. To cope with this problem, this paper introduces a cooperative game theory based framework to evaluate the *power* of each feature. The *power* can be served as a metric of the importance of each feature according to the intricate and intrinsic interrelation among features. Then a general filter feature selection scheme is presented based on the introduced framework to handle the feature selection problem. To verify the effectiveness of our method, experimental comparisons with several other existing feature selection methods on fifteen UCI data sets are carried out using four typical classifiers. The results show that the proposed algorithm achieves better results than other methods in most cases.

1. Introduction

Feature selection, also known as variable selection, is one of the fundamental problems in the fields of machine learning, pattern recognition and statistics. With the new emergences in computer applications, such as social networks clustering, gene expression array analysis and combinatorial chemistry, datasets with tens or hundreds of thousands of features are available. Nevertheless, most of the features in huge dataset are irrelevant or redundant, which lead learning algorithms to low efficiency and over-fitting. Thus, feature selection becomes one of the most active research areas to address this problem. The essential idea of feature selection is to eliminate the irrelevant and redundant features from data set as many as possible. Feature selection in machine learning has been well studied, aiming at finding a good feature subset which produces higher classification accuracy [1]. Also, it is helpful to acquire a better understanding of relationships among the features. Recently several researches have combined feature selection and classification together in various application area to improve the performance of machine learning, e.g., video semantic detection [2], text categorization [3], bioinformatics [4, 5] and intrusion detection [6].

Up to present, several different approaches are employed in feature selection, such as genetic algorithm [7], simulated annealing [8], SVM [9] and boosting method [10]. Furthermore, all of these feature selection methods typically fall into three categories: embedded, wrapper and filter methods. Embedded methods are embedded in and specific to a given machine learning algorithm, and select the features through the process of generating the classifier. Wrappers, evaluating each subset by specified learning algorithms which were treated as a black box, can choose optimal features to yield high prediction performance. One drawback of the wrapper methods, however, is their less generalization of the selected features on other classifiers and high computational complexity in learning, because they are tightly coupled with specified learning algorithms. What's more, they may have a risk of over fitting to the algorithm. Consequently, wrapper methods can hardly deal with large scale problems.

Filter methods are independent of learning algorithms. Instead, they rely on statistical tests over the original features of the training data. In practice, the filter methods have much lower computational complexity than the wrappers, meanwhile, they achieve comparable classification accuracy for most classifiers. Thus, the filter methods are very popular to high-dimension data set. To date, a modest number of efficient filter selection algorithms have been proposed in the literature. It is noteworthy that among various evaluation criterions, Information-theoretic based measurements achieve excellent performance and have drawn more and more attention. This is due to that such measurements capture both linear and nonlinear dependencies without requiring a theoretical probability

* Corresponding author at: College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China, Tel.: +86 0431 85159419; fax: +86 0431 85168337.
E-mail addresses: sunxin1984@yahoo.com.cn (X. Sun), lyh_lb_lk@yahoo.com.cn (Y. Liu).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2012.02.001

distribution or specific model of dependency. However, most of these selectors discard features which are highly correlated to the selected ones although relevant to the target class, which is likely to ignore features which as a group have strong discriminatory power but are weak as individuals [11]. To untie this knot, this paper introduces a cooperative game theory based framework to evaluate the power of each feature. Then a general filter feature selection scheme is presented based on the introduced framework to handle the feature selection problem using any Information-theoretic criteria.

The rest of this paper is structured as follows: In Section 2, related works are briefly reviewed. Section 3 introduces some basic concepts of information theory in feature selection and the necessary background of cooperative game theory. Section 4 provides a cooperative game theory-based framework to evaluate the *power* of each feature, and presents a general filter feature selection algorithm. Section 5 gives experimental results on UCI data sets to evaluate the effectiveness of our approach and some discussions. Conclusions and future work are presented in Section 6.

2. Related work

So far, researchers have proposed lots of selection algorithms to find the optimal feature subset from high-dimension features space [12]. Wrapper method searches for an optimal feature subset tailored to a particular algorithm and a domain [13], e.g., Kabir et al. [14] proposed a wrapper method using neural networks, Inza et al. [15] presented a wrapper method by Estimation of Bayesian Network Algorithm. Embedded methods have better computational complexity than wrapper methods [4]. Guyon et al. [16] propose a embedded method(SVM-RFE) utilizing Support Vector Machine methods based on Recursive Feature Elimination. It has been successfully applied in the area of gene expression analysis. For more detailed reviews on embedded and wrapper methods, readers can refer to the previous literature [11, 13, 17–20] for more information. Since our proposed selection method is independent of any learning algorithms, we focus our attention only on filters. In the following, the state-of-the-art filter methods are briefly reviewed.

For feature selection, one of the most critical challenges is to measure the goodness of a feature subset in determining an optimal one [20]. Unlike wrappers, filters do not employ a learning algorithm to evaluate the selected attribute subsets. Instead, they evaluate the significance of features according to some measurements, such as distance [21–23], rough set theory [24], χ^2 [25], information theory [26] and others. Among the distance based measures, Relief, which is firstly proposed by Kira [21] and later enhanced to support multi-class datasets [22], is one of the most successful ones and adopted Euclidean distance to assign a relevance weight to each feature. The key idea of Relief is to iteratively estimate feature weights according to their ability to discriminate between instances that are near to each other. Having seen abroad spectrum of successful uses of Relief algorithm, Robnik-Šikonja and Kononenko [27] theoretically and empirically investigated and discussed several variations of Relief. Since Relief randomly picks out an instance from training dataset, the optimal results of Relief are not guaranteed. Liu et al. [28] applied selective sampling to Relief in order to obtain results that are better than using random sampling and similar to the results using all the instances. To overcome the disadvantage that Relief lacks a mechanism to deal with outlier data, Sun [29] proposed an iterative Relief algorithm to alleviate the deficiencies of Relief by exploring the framework of the Expectation-Maximization algorithm. Other distance based measures, such as Kolmogorov Distance and Normalized Compression Distance, are also popularly used in feature selection [23]. Hu et al. [30] introduced a concept of neighborhood margin and neighborhood soft margin to measure the minimal distance between different classes. They utilized the criterion of neighborhood soft margin to evaluate the quality of candidate features and construct a forward greedy algorithm for feature selection.

Rough set theory has been proven to be an efficient tool for modeling and reasoning with uncertainty information. Feature selection under rough set theory is a consistency-based method [31], which attempts to retain the discriminatory power of original features for the objects from the universe [32]. Recent years, researchers have focused their attention on feature selection algorithms based on rough sets [32, 33]. However, algorithms based on rough sets are often computationally time consuming. Qian et al. [34] introduced a theoretic framework based on rough set theory, which is called positive approximation and can be used to accelerate a heuristic process for feature selection from incomplete data. Based on this framework, they also presented a general heuristic incomplete feature selection algorithm as an application of the proposed accelerator. Recently, another consistency-based method [35] have been proposed to use pairwise constraints for feature selection by Zhang et al. They devised two novel score functions based on pairwise constraints to evaluate the feature goodness and named the corresponding algorithms as Constraint Score.

The prediction capability of individual feature and the intercorrelation of feature subset are two important aspects in feature selection. There exist broadly two approaches to measure the correlation among features. One is based on classical linear correlation and the other is based on information theory [11]. Recent years a large amount of literatures on information theoretic ranking criteria have been proposed. A major advantage of information theoretic criteria is that they capture higher order statistics of the data [26]. Battiti et al. [36] investigated the application of the mutual information criterion to evaluate a set of candidate features and to select an informative subset to be used as input data for a neural network classifier. Then, an algorithm MIFS was proposed that takes both the mutual information with respect to the output class and with respect to the already selected features into account. However, the MIFS algorithm may fail when redundant features have much information about the output. Nojun et al. [37] proposed an improved algorithm of feature selection that makes more careful use of the mutual information between input attributes and others than the original MIFS. Kwak and Choi [38] proposed a new method of calculating mutual information between input and class variables based on the Parzen window, and applied this to a feature selection algorithm for several classification problems. Novovicova et al. [39] proposed a new sequential forward selection algorithm mMIFS-U that uses novel estimation of the conditional mutual information between candidate feature and classes given a subset of already selected features. Because of the difficulty in directly implementing the maximal dependency condition, Peng et al. [40] first derived an equivalent form, called minimal-redundancy-maximal-relevance criterion (mRMR), for first-order incremental feature selection. Then they presented a two-stage feature selection algorithm by combining mRMR and other more sophisticated feature selectors. Yu and Liu [41] introduced a new framework that decouples relevance analysis and redundancy analysis. They developed a correlation-based method named symmetrical uncertainty(SU) for relevance and redundancy analysis, and then removed redundant features by approximate Markov Blanket technique. In traditional selectors, mutual information is estimated on the whole sampling space. This, however, cannot exactly represent the relevance among features. To cope with this

problem, Liu et al. [42] proposed a feature selection algorithm based on dynamic mutual information, which is only estimated on unlabeled instances.

Hu et al. [43] reformed Shannon's information measures into relation matrix representations to measure the information quantity in a set of fuzzy sets and applied the proposed measure to calculate the uncertainty in fuzzy approximation spaces. Lee et al. [44] presented an efficient classifier with feature selection based on fuzzy entropy for pattern classification, and Shie et al. [45] presented a new method for feature subset selection which can deal with both numeric and nominal features based on the proposed fuzzy entropy measure for handling classification problems.

Information gain is another common measure of association currently used in data mining. Lee [46] proposed an information gain and divergence-based feature selection method which strives to reduce redundancy among features while maintaining information gain in selecting appropriate features.

This section can not be a complete review of all algorithms and just briefly describes the most related studies. Yet we hope that the description can help to understand our studies better and guide the readers in interesting research directions.

3. Preliminaries

In this section, the formalism of feature selection is presented, then common concepts about information theory [47] and cooperative game theory are introduced.

3.1. Feature selection

Throughout the paper, the following notations are used. Formally, assume that $T=(F, O, C)$ is a sampling dataset, where $F=\{f_1,\ldots, f_m\}$, $O=\{o_1,\ldots, o_n\}$ and $C=\{c_1,\ldots, c_k\}$ are the sets of features, instances and target classes, respectively. For each instance $o_i\in O$, it can be represented as a value vector of features, i.e., $o_i=(v_{i1},\ldots, v_{im})$, where v_{ij} is the value of feature f_j corresponding to the sample o_i.

The main goal of feature selection is to find the minimum subset S of maximal relevant features from set F. Meanwhile the subset S is optimized for the performance of machine learning algorithm. Obviously, the rigorous way to obtain the subset S is exhaustive search and evaluation of all the possible subsets of F, which has been demonstrated as *NP*-Completeness [48]. To circumvent this problem, many heuristic subset search or selection strategies have been proposed [19].

3.2. Information theory

For the sake of convenience, we only deal with finite random variables with discrete values. Let X be a discrete random variable and probability density function $p(x)=Pr\{X=x\}$. The entropy $H(X)$ of a discrete random variable X is defined by

$$H(X)=-\sum_{x\in X}p(x)\log p(x). \tag{1}$$

Mutual information (MI) is a measure of the amount of information shared by two variables X and Y. Consider two random variables X and Y, the mutual information $I(X;Y)$ is defined as

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \tag{2}$$

The conditional mutual information $I(X;Y|Z)$ is defined as the amount of information shared by variables X and Y, when Z is given. It is formally defined by

$$I(X;Y|Z)=\sum_{x\in S_X}\sum_{y\in S_Y}\sum_{z\in S_Z}p(x,y,z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)}. \tag{3}$$

In this paper, conditional mutual information is used to evaluate the impact of newly emerging variable on the information shared by variables with the class.

3.3. Basic concepts for cooperative game theory

Coalitions: $N=\{1, 2,\ldots, n\}$ is the set of players. A coalition [49] is a subset of players, N, which is able to make a binding agreement.

A coalitional game consists of a finite set of players N, and a characteristic function. The characteristic function is a scalar valued function, $v(K)$, that associates $v(K)\in R$ with each $K\subset N$. The characteristic function value for the empty coalition is 0. That is $v(\emptyset)=0$. The solution of cooperative games represents the contribution of each player to the game by constructing a value function, which assigns a value to each player. Banzhaf power index [50] was proposed by Banzhaf, which yields a unique outcome in coalitional games, to measure the contribution of players in the game. It is based on counting, for each player, the number of coalitions to which the player is crucial to winning.

4. Feature evaluation and selection with coalitional game theory

4.1. Problem statement

For most feature selection algorithms based on information theory, feature (or subset) that has high relevance with the class and low redundancy among selected features will be selected in each iteration. The major disadvantage of these algorithms is that they disregard the dependencies between the candidate feature and unselected features. Consequently, interdependent features, weak as individuals but having strong discriminatory power as a group, are likely to be ignored. Guyon et al. [18] also constructed an example by the famous XOR problem to illuminate that two variables which are useless by themselves can be useful together.

To illustrate the problem clearly, a simply sample is given in Table 1 to select not more than three features. Obviously, it can be seen that subsets containing feature f_2 and f_3 exhibit the best performance. However, most of the famous feature selection algorithms, including BIF, MIFS, mRMR and CMIM, regard *subset* (f_4, f_1, f_2) as optimal. The main reason for this disadvantage is that Information-theoretic based measurements disregard the intrinsic structure [51] among features. To untie the knot, a practicable method is needed to retain the useful intrinsic structures for feature selection.

Table 1
Sample $T=(F, O, C)$.

	f_1	f_2	f_3	f_4	Class	f_4	$f_{1,4}$	$f_{2,3}$	$f_{2,4}$	$f_{3,4}$	$f_{1,2,3}$	$f_{1,2,4}$	$f_{1,3,4}$	$f_{2,3,4}$
o_1	1	2	1	3	I	√	√	√	√	√	√	√	√	√
o_2	2	2	2	1	III									
o_3	2	1	3	2	II			√			√			√
o_4	3	1	3	1	II			√			√			√
o_5	2	1	1	2	III									
o_6	3	2	1	3	I	√	√	√	√	√	√	√	√	√
o_7	2	3	3	1	III									

4.2. Class-based redundancy, interdependence and independence analysis

In most previous literature, candidate features which are highly correlated with the selected features will be regarded as redundancy and discarded, e.g., mRMR introduced the criterion namely "Min-Redundancy" to eliminate the redundant features. However, it is likely to disregard the intrinsic interdependent groups which as a group have strong discriminatory power but are weak as individuals. The main reason is that features which have been labeled "redundancy" may not be real redundancy. In this work, information theoretic measurements including MI and CMI are adapted to distinguish the relationship of redundancy, interdependence and independence between features.

(1) Redundancy
A feature is said to be redundant if one or more of the other features are highly correlated with it, and its relevance with the target class can be reduced by the knowledge of any one of these features. That is, if a feature f_j is real redundant with f_i, then the relevance between f_j and target class can be reduced under the condition of f_i. The formulation is defined as follow:

$$I(f_j; class|f_i) < I(f_j; class). \tag{4}$$

(2) Interdependence
Interdependence implies each member in the relationship cannot function apart from one another, namely the impact of each feature on the classification performance can not be ignored and replaced. Thus, suppose f_i and f_j are interdependent on each other, then the relevance between f_j and target class can be increased conditioned by f_i. Namely, two features f_i and f_j are interdependent on each other if the following form is satisfied.

$$I(f_j; class|f_i) > I(f_j; class) \tag{5}$$

(3) Independence
If two features f_i and f_j are completely independent (or irrelevant), then the relevance between target class and any one of them will not be changed by the other emerging as a condition. That is

$$I(f_j; class|f_i) = I(f_j; class). \tag{6}$$

It is easy to see that the optimal feature subset is the one in which all the features are relevant to the target class and interdependent on each other.

4.3. Feature evaluation framework based on cooperative game theory

In natural large-scale data sets, there are intrinsic correlations among variables [51], such as causality, interdependence and unidirectional dependent. The existence of intrinsic correlative structures among variables results in different importance of every individual. Our contributions focus on evaluating the importance (or power) of each feature using the Banzhaf power index and retaining the useful intrinsic structures for feature selection.

The original definition of Banzhaf power index is described as follows [50]: A winning coalition is one for which $v(S)=1$ and a losing coalition is one for which $v(S)=0$. Each coalition $S\cup\{i\}$ that wins when S loses is called a swing for player i, because the membership of player i in the coalition is crucial to the coalition winning. Let $\sigma_i(N, v)$ be the number of swings for i, and let $\sigma_o(N, v)=\sum_{i\in N}\sigma_i(N,v)$ be the total number of swings of all players in the game. Then the normalized Banzhaf index is $b_i(N,v)=\sigma_i(N,v)/\sigma_o(N,v)$. The generalization is made by using the formula

$$b_i(N,v) = \frac{1}{2^{n-1}} \sum_{S\subseteq N\setminus i} \Delta_i(S). \tag{7}$$

where $\Delta_i(S)$ is the marginal contribution of player i and defined as

$$\Delta_i(S) = v(S\cup i) - v(S). \tag{8}$$

Banzhaf power index has been applied to problems like (a) distribution of power in UN Security Council, (b) to understand the Electoral College method of electing US presidents. Some famous examples in the literature [52] can lead readers to a better understanding of the application and calculation of Banzhaf power index.

Banzhaf power index measures the distribution of power among the players in the voting game, which can be transformed into the arena of feature selection attempting to estimate the *power* of each feature. The idea is motivated by the observation that every subset of features can be regarded as a candidate subset for the final selected optimal subset, thus, the power of each feature can be measured by averaging the contributions that it makes to each of the subset which it belongs to.

Let coalition K be a candidate subset and feature $f_i(f_i\notin K)$ is to be estimated. Then the impact of the feature f_i on coalition K can be evaluated as the following description: Let $\eta_i(K)$ be the number of features of which are fall into interdependence relationship with the feature f_i, then $\mu_i(K)$ be the number of redundant or independent with the feature f_i,. The impact of feature f_i on subset K is evaluated based on the ratio $p=\mu_i(K)/\eta_i(K)$. For the sake of convenience, a threshold value τ is defined before the following discussion. We named the coalition $K\cup\{i\}$ lost if $p<\tau$, which means the coalition K is unstable when f_i is emerging, because more than $(1-\tau)$ percentage of features reduced their relevance with target class. On the contrary, conditions of $p\geq\tau$ imply that coalition K can exhibit better performance if f_i is added, in which coalition $K\cup\{i\}$ is termed win.

Based on the above discussions, we redefine the marginal contribution $\Delta_i(S)$ in the context of feature selection as

$$\Delta_i(S) = \begin{cases} 1 & p\geq\tau \\ 0 & p<\tau \end{cases} \tag{9}$$

which means coalitions for each feature f_i crucial to winning are defined as ones exhibit better performance when the feature f_i is joining. According to majority principle that the majority can control the coalition, we assign the threshold value as 1/2 which means the given feature will win the coalition if more than half of the features are interdependent with the given feature.

Due to the fact that every coalition is possible to be a subset of the final result, it is an effective approach to evaluate the impact of feature f_i by calculating the proportion of winning coalitions under conditions of f_i to all possible coalitions.

Algorithm 1. Feature evaluation framework based on cooperative game theory.

Input: A training sample O with feature space F and the target C.
Output: Pv: Banzhaf power index vector of F.

1. $Pv=0$, limit value ω and threshold value τ are assigned;
2. **For** each feature $i\in F$ **do**
3. Create coalitions set$\{\pi_1,\ldots,\pi_t\}$ over $F\setminus i$ limited by ω;
4. **For** each $\pi_j\in\{\pi_1,\ldots,\pi_t\}$ **do**
5. Calculate payoff function $\Delta_i(\pi_j)$ using formula (9);
6. End

7. Calculate the Banzhaf power index $b_i(N,v)$ using formula (7);
8. $Pv(i)=b_i(N,v)$;
9. End
10. Normalized the vector Pv.

More specifically, details of the feature evaluation framework based on cooperative game theory are presented in Algorithm 1. The output of this evaluation framework is a vector Pv of which each element $Pv(i)$ represents the normalized Banzhaf power index of feature f_i. It is noticed that Banzhaf power index is only a metric estimating the importance of every feature based on the intrinsic correlative structures among features. Thus, to select features based on our evaluation framework, a metric reflecting the feature's relevance to target class and a heuristic search strategy are also needed. In the following section, a general feature selection scheme of combining our evaluation framework with any information theoretic criteria is presented.

4.4. A general feature selection scheme

Algorithm 2. CoFS: A general feature selection scheme with cooperative game

Input: A sampling dataset $T=(F, O, C)$ and Pv.
Output: Selected feature subset S.

1) Initialize parameters: $S=\emptyset$, $F=\{f_1,\ldots,f_m\}$, $k=0$;
2) While $k<\delta$ do
3) **For** each feature $f_i\in F$ **do**
4) Calculate its value of criterion $J(f_i)$;
5) Calculate its victory criterion $V(f_i)=J(f_i)\times Pv(i)$;
6) End
7) Choose the feature f_i with the largest $V(f_i)$;
8) $F=F\backslash\{f_i\}$, $S=S\cup\{f_i\}$;
9) $k=k+1$;
10) End

Features can be selected in many different ways, such as forward, backward and random [17]. In this paper, we consider the feature selection procedure in a straight forward way. Features will be weighted in each iteration by most of the traditional Information-theoretic based selectors according to its own criterion. On this basis, we reweight the features according to its impact on intrinsic correlative structures among features, based on the evaluation results of our framework. For example, the weight of features with higher(or lower) Banzhaf power index will be raised(or reduced). Without loss of generality, to handle the feature selection problem, a general scheme is presented based on our evaluation framework by employing any criteria, such as BIF, MIFS and mRMR. The pseudo codes are outlined as Algorithm 2. This selection procedure will be terminated if the number of selected features is larger then the user-specified threshold δ.

In order to select the optimal feature in each iteration, victory criterion $V(f)$ is defined to evaluate the superiority of each feature over others. Criterion function $J(S)$ is used to pick out the feature f_i which has high relevance with the target class and low redundancy among subset $S\cup\{f_i\}$ from remaining features. However, as mentioned above, the feature picked out by $J(S)$ is not really the optimal one. Therefore, we utilize the Banzhaf power index $Pv(i)$, which denotes the inherent impact of feature f_i on the whole feature space, to regulate the relative importance of its evaluation value $J(f_i)$ for feature selection. This general scheme gives an effective way to overcome the disadvantage that traditional information-theoretic based selectors is likely to ignore the intrinsic correlative structures among variables. There exist so many information-theoretic based feature selection methods that can be employed as the criterion function in the proposed general feature selection scheme, such as BIF, MIFS and mRMR. In this work, as an illustration, mRMR is employed as the criterion function.

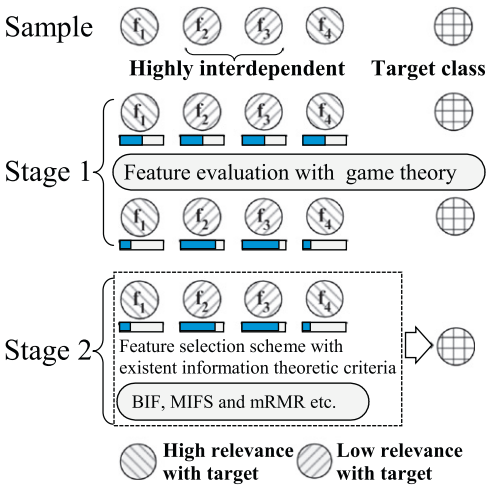

Fig. 1. A simple illustration for feature selection process of our method.

In the following we extend the simple toy example in Table 1 to show the process of our method(shown in Fig. 1).

Stage 1: We firstly introduce a framework based on cooperative game theory to evaluate the contributions that the features make. As an example, we calculate the Banzhaf power index of f_2 in the condition of $\omega=2$: the total number of coalitions(subsets) of feature set $F\backslash i$ is $|\Pi_\omega|=6$; and the winning coalitions of f_2 are $\{f_3\}$, $\{f_1, f_3\}$, and $\{f_3, f_4\}$ where more than half of the menbers are interdependent with feature f_2; then Banzhaf power index of f_2 is $b_2(N,v)=1/|\Pi_\omega|\sum_{S\subseteq\Pi_\omega} v(S\cup i)=0.5$; similarly, the values of Banzhaf power index for f_1, f_3 and f_4 are 0, 0.5, 0 (The Banzhaf power index hardly equals zero in complicated real data sets). As shown in Fig. 1, the features which fall into a interdependent relationship are assigned a high weight according to the normalized Banzhaf power index.

Stage 2: Then a general feature selection scheme is proposed to handle the feature selection problem by using a re-weighting mechanism for any existent information theoretic criteria.

4.5. Computational complexity reduction

Theoretically, the calculation of the Banzhaf power index requires summing over all possible subsets of features, which can extremely increase the computational complexity. In fact, the number of features correlated with a certain feature is much smaller than the total number of features in the real data set. It is unnecessary to consider all coalitions for features, especially large coalitions. Moreover, the consideration of large coalitions may deteriorate the selector's performance, for the probability that the selected optimal subset contains whole features of large coalitions is very low. Thus, we suggest a limit value ω being a bound on the coalition size. The formula (7) can be redefined as

$$b_i(N,v)=\frac{1}{|\Pi_\omega|}\sum_{S\subseteq\Pi_\omega} v(S\cup i), \tag{10}$$

where Π_ω is the set of subsets of feature set $F\backslash i$ limited by ω.

Moreover, in our proposed approach, whether a coalition win or not is pivoted on the number of features increasing (or reducing) its relevance with target class when the condition is given. So, first the number of winning coalitions having only one member(denoted as W_1) can be calculated with time complexity $O(n)$. Then, the number of winning coalitions having more than one member can be calculated based on W_1 by the knowledge of combinatorial mathematics and dynamic programming technique. Dynamic programming' is an efficient programming technique for solving certain combinatorial problems[53]. For instance, W_2 can be calculated as $W_2 = C^2_{W_1} + C^1_{W_1} \cdot C^1_{n-W_1}$. In this way, the problem can be solved with low computational complexity.

5. Experiments and results

Before discussing the results of experiments, a brief introduction of data sets and experimental setup are given first.

5.1. Data sets and experimental setup

To test the proposed method empirically, fifteen real world data sets from UCI's machine learning data repository [54] are adopted in our simulation experiments, as shown in Table 2. These datasets contain various numbers of features and come from different domains, such as biology, computer, life and physical. However, documentations for these datasets (available online at the UCI Machine Learning data archive) state that there exist some missing values arisen from various aspects. We replace each missing value with the mean for numeric attributes and the mode for nominal ones [55]. For the continuous features, it is difficult to compute their information entropies using a limited number of instances. For convenience, we discretize continuous-valued attributes into multiple intervals using a supervised discretization method named MDL method [56]. Discretization of attributes can reduce the learning complexity and help to understand the dependence between attributes and the target.

In addition to feature selection algorithms, we employed four representative classifiers, i.e., Naive Bayes, SVM, 1-Nearest Neighbor and C4.5, which are the most influential algorithms that have been widely used in the data mining community [57]. The experimental workbench is Weka (Waikato environment for knowledge analysis), which is a collection of machine learning algorithms for data mining tasks. The parameters of classifiers for each experiments are set to default values of Weka. For estimating the performance of classification algorithms, *10*-fold cross-validation is used. In *10*-fold cross-validation the data is first partitioned into*10* nearly equally sized folds. Subsequently *ten* iterations of training and validation are performed such that within each iteration a different fold of the data is held-out for validation while the remaining *nine* folds are used for learning. To determine whether the experimental results are significant or not, paired *t*-tests between accuracies with CoFS and with other selectors at a time have been carried out. Throughout this paper, the difference of accuracies is considered significantly different if its *p*-value is less than 0.05 (i.e., confidence level greater than 95%) according to a paired *t*-test.

Table 2
Summary of datasets in our experiments.

No.	Dataset	Samples	Features	Classes
1	Glass	214	9	7
2	Pen recognition	10992'	16	10
3	Lymphography	148	18	4
4	Cardiotocography	2126	22	3
5	Hypothyroid	3772	29	4
6	Spectf	267	44	2
7	Multi-feature zer	2000	47	10
8	Synthetic	600	60	6
9	Molecular splice	3190	60	3
10	Optical recognition	5620	65	10
11	Musk (Version 2)	6598	166	2
12	Multi-feature factors	2000	216	10
13	Multi-feature pixel	2000	240	10
14	Arrhythmia	452	279	16
15	Isolet	1559	618	26

We empirically evaluate the performance of our proposed method by comparing with three typical feature selectors: mRMR, ReliefF and IG. The mRMR [40] criterion is employed in the proposed scheme as an illustration in this work. Thus, it is essential to exhibit the performance comparison between mRMR filter and our proposed method, in order to verify whether the proposed scheme can improve the performance of original information-theoretic based feature selection algorithm. ReliefF [22] is a popular instance-based feature weighting algorithms and Information Gain (IG) measures the decrease in entropy when the feature is given vs. absent [46]. They are all well-known and have excellent performance. For ReliefF, we use 5 neighbors and 30 instances throughout the experiments as suggested by Robnik-Sikonja and Kononenko [27], which is also used in the literature [41]. The parameter ω in our feature evaluation method (CoFS) is set to 3.

5.2. Feature selection and classification results.

The effectiveness of a feature selection algorithm can be simply and directly measured by the classification performance on different datasets for classifiers. Since the arrangement of features ranked by these four selectors is different, we compare the top classification accuracies across different "acceptable" numbers of selected features on the real datasets. The "acceptable" number means that about a third of original features remained for low-dimension datasets (less than 100) and at most thirty features for high-dimension datasets (more than 100).

Table 3 records the classification accuracies on fifteen real datasets for classifiers using four feature selection algorithms, in which the "UnSelect" column records the accuracies of classifiers on datasets with original features. Additionally, the bold value means that it is the largest one among these four feature selectors under the same classifier. Notation "$\circ$" (or "**v**") represents the value of current entry is significantly worse (or better) than the corresponding one in the "CoFS" column in statistical *t*-test. The average value of accuracies with the same selector is given in the row labeled as "Ave.". The average number of features selected by each feature selection algorithm is recorded in the "AvF" row. In order to investigate the efficiency of selectors with the same classifier, we give a new definition named "Average efficiency value" as below.

$$AEV = \frac{Ave.}{AvF} \tag{11}$$

The "Average efficiency value" is exhibit in the "*AEV*" row for different selectors(including the original features)in the same classifier.

We first make a comparison between CoFS and mRMR so as to verify whether the original information-theoretic based selection algorithm is improved by the proposed evaluation framework. The results in Table 3 show that the performance using CoFS are better than those of mRMR. For example, there are eleven cases that the classification accuracies of CoFS are higher than those of mRMR in the NB classifier. And the average value of classification accuracy (in the "Ave." row) for the fifteen real datasets also denotes that our method perform better than mRMR, meanwhile,

Table 3
The comparison of classification accuracies of classifiers on fifteen datasets (the corresponding number of selected features in parentheses).

No.	Naive Bayes					SVM				
	UnSelect	CoFS	mRMR	ReliefF	IG	UnSelect	CoFS	mRMR	ReliefF	IG
1	74.29%	**73.60%**(6)	71.88%(5)°	71.64%(5)°	71.88%(5)°	66.82%	67.81%(4)	**69.27%**(4)	**69.27%**(4)	**69.27%**(4)
2	87.90%	**83.62%**(6)	82.22%(6)°	80.70%(6)°	68.19%(6)°	99.40%	**94.10%**(6)	93.88%(6)°	92.81%(6)°	85.88%(6)°
3	83.78%	**84.38%**(6)	80.64%(5)°	78.47%(4)°	79.13%(4)°	79.05%	**79.28%**(6)	75.69%(6)°	79.25%(5)	79.25%(5)
4	92.47%	**98.73%**(6)	98.54%(6)	98.40%(2)	98.40%(2)	98.44%	**98.78%**(3)	98.40%(3)°	98.68%(5)	98.50%(4)
5	98.54%	**98.78%**(5)	**98.78%**(5)	98.57%(9)	98.70%(6)	98.67%	**99.34%**(6)	**99.34%**(6)	99.15%(9)	**99.34%**(6)
6	79.55%	**85.91%**(5)	**85.91%**(5)	81.77%(7)	80.30%(5)	79.55%	**87.02%**(9)	86.61%(9)	85.50%(18)	85.48%(16)
7	74.30%	**69.95%**(13)	66.90%(13)°	68.10%(13)	63.70%(13)°	81.40%	**78.25%**(14)	75.95%(14)°	77.35%(14)	72.80%(14)°
8	98.33%	**93.33%**(9)	91.50%(17)°	82.67%(17)°	80.60%(7)°	99.50%	**96.00%**(12)	94.67%(20)	86.83%(15)°	85.17%(14)°
9	95.36%	**95.55%**(13)	95.14%(13)°	93.57%(13)°	95.08%(13)	92.45%	**93.98%**(7)	**93.98%**(7)	91.35%(14)°	**93.98%**(7)
10	92.31%	**91.44%**(19)	91.28%(18)	89.27%(20)°	88.93%(15)°	98.93%	**98.26%**(20)	97.65%(20)°	97.35%(20)°	97.26%(20)°
11	91.47%	92.32%(19)	**92.51%**(19)	89.00%(7)°	92.03%(13)	77.16%	**95.47%**(26)	95.21%(26)°	94.92%(7)	94.80%(15)°
12	93.65%	**92.00%**(17)	91.10%(17)°	83.85%(30)°	88.35%(28)°	97.65%	**96.70%**(19)	95.75%(19)°	94.95%(30)°	94.90%(26)°
13	93.30%	**81.20%**(9)	79.55%(10)°	72.95%(17)°	62.20%(20)°	94.80%	**90.30%**(20)	88.30%(20)°	81.95%(20)°	69.45%(20)°
14	75.00%	76.79%(18)	**78.10%**(29)	75.00%(19)	74.34%(17)°	59.51%	**75.66%**(18)	75.00%(24)°	73.68%(18)°	73.45%(23)°
15	89.03%	**76.65%**(29)	69.79%(29)°	56.90%(30)°	51.44%(30)°	93.97%	**83.58%**(30)	76.20%(30)°	64.34%(30)°	62.99%(30)°
Ave.	87.95%	**86.28%**	84.92%	81.39%	79.55%	87.82%	**88.97%**	87.73%	85.82%	84.17%
AvF	126	**12**	13	12	12	126	**13**	14	14	14
AEV	0.70%	**7.19**%	6.47%	6.13%	6.49%	0.70%	**6.35**%	6.15%	5.99%	6.01%
	1-Nearest neighbor					C4.5				
	UnSelect	CoFS	mRMR	ReliefF	IG	UnSelect	CoFS	mRMR	ReliefF	IG
1	79.90%	**78.49%**(4)	69.92%(4)°	69.92%(4)°	69.92%(4)°	73.83%	**70.58%**(3)	66.42%(3)°	66.42%(3)°	68.34%(3)°
2	96.72%	**89.35%**(6)	88.58%(6)	88.57%(6)	81.84%(6)°	88.55%	**87.41%**(6)	86.20%(6)°	86.33%(6)°	79.90%(6)°
3	83.78%	79.95%(6)	81.95%(6)	78.00%(6)	**82.23%**(6)	78.37%	**80.69%**(7)	74.97%(6)°	77.26%(6)°	79.73%(3)
4	96.09%	**98.68%**(3)	98.35%(3)°	98.21%(3)°	98.12%(3)°	98.77%	**98.97%**(5)	98.59%(5)°	98.82%(5)°	98.73%(4)
5	97.91%	**99.15%**(5)	**99.15%**(5)	97.96%(3)°	98.99%(6)	99.31%	**99.15%**(5)	**99.15%**(5)	98.12%(3)	**99.15%**(6)
6	82.89%	**85.91%**(5)	**85.91%**(5)	85.51%(9)	85.51%(16)	82.52%	85.91%(5)	85.91%(5)	**86.28%**(14)	84.37%(16)
7	71.20%	**68.30%**(13)	64.95%(13)°	67.30%(13)	60.25%(13)°	64.65%	64.20%(9)	61.00%(8)°	**66.00%**(15)	60.80%(12)°
8	98.16%	**92.00%**(8)	89.50%(18)°	86.17%(17)°	83.00%(18)°	92.33%	**90.50%**(8)	84.67%(11)°	82.67%(13)°	79.33%(7)°
9	74.67%	**90.25%**(5)	**90.25%**(5)	82.88%(5)°	**90.25%**(5)	94.39%	**94.36%**(7)	**94.36%**(7)	93.07%(12)°	**94.36%**(7)
10	94.12%	**92.47%**(20)	91.32%(20)°	90.16%(20)°	90.52%(20)°	77.93%	**79.75%**(16)	78.74%(13)°	78.67%(20)°	78.51%(13)°
11	90.89%	95.71%(25)	95.76%(25)	**96.01%**(19)	95.89%(24)	91.01%	95.94%(26)	**96.21%**(30)	94.36%(20)°	95.30%(29)°
12	95.60%	**92.40%**(17)	92.15%(30)	86.50%(30)°	87.65%(24)°	82.35%	81.65%(16)	**82.60%**(21)	78.50%(24)°	81.50%(19)
13	96.15%	**85.55%**(16)	83.60%(16)°	77.40%(17)°	62.45%(17)°	78.65%	**78.30%**(18)	76.60% (18)°	74.65%(26)°	68.85%(26)°
14	68.80%	71.23%(22)	**71.88%**(17)	71.20%(29)	71.45%(12)	71.23%	**75.69%**(19)	75.22%(20)°	75.44%(30)	74.56%(28)°
15	85.12%	**66.52%**(28)	64.98%(28)°	54.65%(21)°	53.62%(30)°	73.38%	55.49%(19)	**62.28%**(19)**v**	55.74%(30)	57.41%(30)
Ave.	87.47%	**85.73%**	84.55%	82.03%	80.78%	83.15%	**82.57%**	81.53%	80.82%	80.06%
AvF	126	**12**	13	13	14	126	**11**	12	15	14
AEV	0.69%	**7.03**%	6.31%	6.09%	5.94%	0.66%	**7.33**%	6.91%	5.34%	5.75%

the "*AEV*" row indicates the CoFS is higher efficiency. For the rest of the three classifiers, one may also observe that the CoFS clearly surpasses others in most cases. Among the methods, we can see that the CoFS approach achieves significantly higher classification accuracy than the other methods in most cases (*paired t-test with* 95*% confidence*). For example, the numbers of cases for which CoFS achieves significantly higher classification accuracy over mRMR, ReliefF and IG are ten, eight and nine out of fifteen cases in the SVM classifier, respectively. We also make a comparison at the aspect of win/tie/loss and the results are given in Table 4. It can be seen from this table that CoFS outperforms mRMR selection method in most cases. Overall, the proposed CoFS approach exhibits better performance than the origin selection algorithm employed as the criterion function, and gives an effective way to find the optimal feature subset.

In addition, the performances using ReliefF and IG are also shown in Table 3. We can see that all these methods achieve significant reduction of dimensionality by selecting only a small portion of the original features. And among these methods, CoFS on average selects the smallest number of features but achieves the highest classification accuracy. As an illustration, for CoFS, the average efficiency in the four classifiers are 7.19%, 6.35%, 7.03%, and 7.33%, respectively, which are much higher than other selectors. Furthermore, CoFS achieves the highest classification performance over three-quarter datasets in all classifiers, which is higher than other selection algorithms. From Table 4, we can see that CoFS performs better than other three feature selection methods in most cases under any classifier. As an example, the entry "14/0/1" between CoFS and ReliefF under condition SVM denotes that CoFS wins on 14 and losses only one case in comparison with ReliefF on classification capability of the SVM classifier. It is noticeable that the proposed method works well in all classifiers.

Table 4
A comparison of win/tie/loss between CoFS and other selectors.

Win/tie/loss	mRMR	ReliefF	IG
CoFS			
Naive Bayes	11/2/2	15/0/0	15/0/0
SVM	12/2/1	14/0/1	12/2/1
1-NN	9/3/3	14/0/1	12/0/3
C4.5	9/3/3	12/0/3	12/2/1

To illustrate the validity of the selected features by our method, we compare the correct classification accuracies with these four feature selectors using different classifiers. The experimental datasets are Lymphography, Synthetic, Multi-feature pixel and Isolet, where their numbers of features differ greatly (range from 18 to 618). It can be noticed from Table 3 that the selected features exhibit inconsistent performances in different classifiers.

For example, our proposed method works well in the Naïve Bayes classifier than others. It has been shown that different learning algorithms need different feature selectors to suit their learning bias [58]. For the sake of impartiality, we calculate the mean value of performance of classifiers for each selector. The comparison results are shown as Fig. 2. The number k on X-axis of Fig. 2 refers to the first k features with selected order by different selectors. The Y-axis represents the mean performance of classifiers of the first k features. The results in Fig. 2 indicate that the mean performance of classifiers with CoFS is superior to others. we can see that all plots of CoFS are higher than those of other methods with an acceptable number of features, although plots with the first few features are lower than others.

In our framework, parameter ω is introduced to improve the accuracy of feature evaluation and reduce the computational complexity. It is assumed as the size of interdependent groups. In fact, the value of the size is unequal according to various realistic data sets, and even unequal within the same dataset. To investigate the impact of limit value ω on the accuracy of feature evaluation, we compare the classification performance of SVM against different value of ω on three data sets (Multi-feature pixel, Arrhythmia and Isolet) as shown in Fig. 3. We can see that our method performs better when the value of $\omega \in [2, 4]$. For example, the classification accuracy reaches its highest when ω is set to 3 for multi-feature pixel and Arrhythmia data sets. For the sake of convenience, we suggest the limit size a moderate value $\omega=3$ in all above experiments.

As previously stated, all feature combinations within limited size ω need to be considered in order to compute the Banzhaf power index. And the proposed general feature selection scheme CoFS combines our cooperative game theory based feature evaluation framework with existent information theoretic criteria. All of these lead to increasing time-complexity for feature selection problem. We have intorduced combinatorial mathematics and dynamic programming technique to reduce the time complexity. Therefore, it is necessary to exhibit whether the evaluation problem can be solved at present in acceptable running time. In this paper, mRMR is employed in the proposed scheme as an illustration. Thus we record the running time for CoFS and mRMR to rank features on eight representative UCI data sets. All experiments are conducted on a Pentium IV, with a CPU clock rate of 2.8GHz and 1G main memory. The proposed algorithm and mRMR have been implemented in Matlab.

We can observe from Table 5 that the proposed feature evaluation framework can evaluate features with acceptable computational complexity.

5.3. Discussions

For datasets with a large number of features, there exist various intrinsic correlations among variables [48]. Advantages of filter selectors are that they are fast and easy to interpret. There are also some disadvantages of using filters [11], such as (i) redundant features may be included and (ii) some features which as a group have strong discriminatory power but are weak as individual features will be ignored. To cope with these problems, a new method for feature evaluation and selection has been proposed in this paper. As is known to all, it is difficult to discover the association relationship among features exactly. We can not guarantee that our method retains all useful interdependent groups or the whole interdependent group, however, the

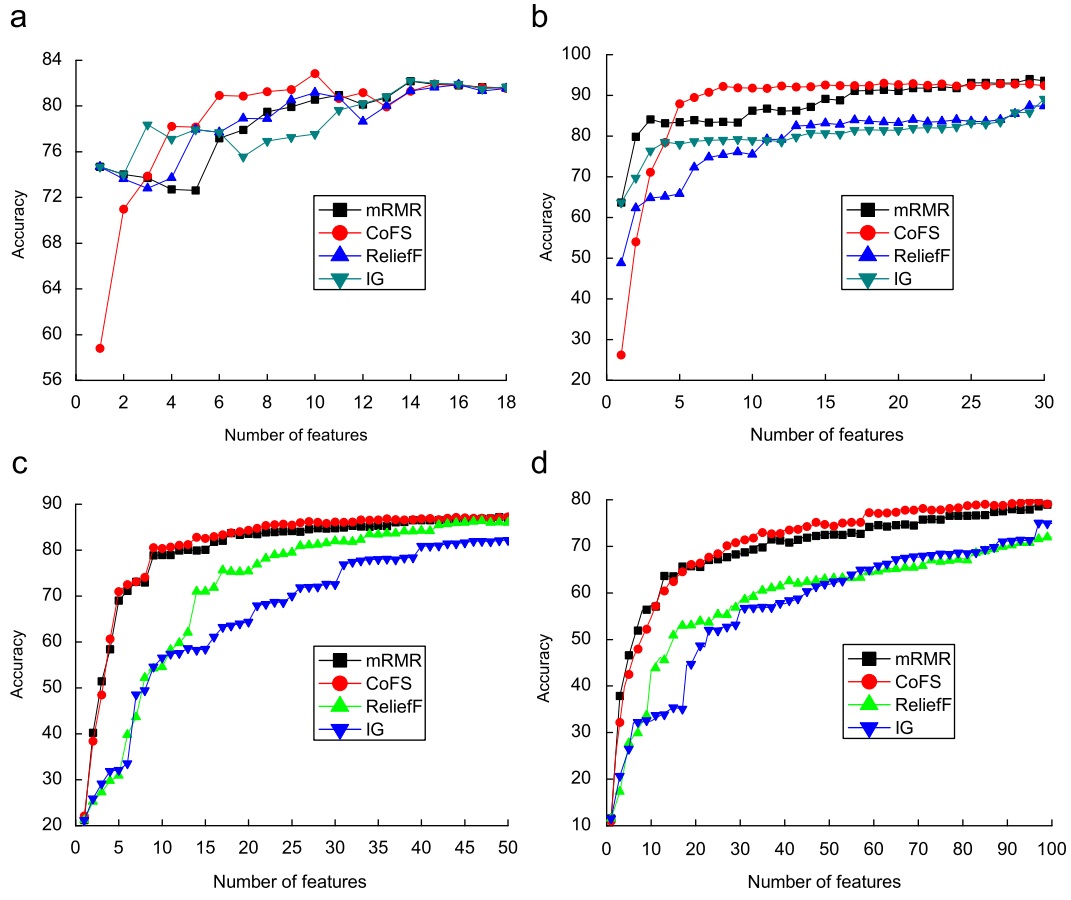

Fig. 2. Accuracies vs. different numbers of selected features on four UCI datasets: (a) lymphography, (b) synthetic, (c) multi-feature pixel, (d) isolet.

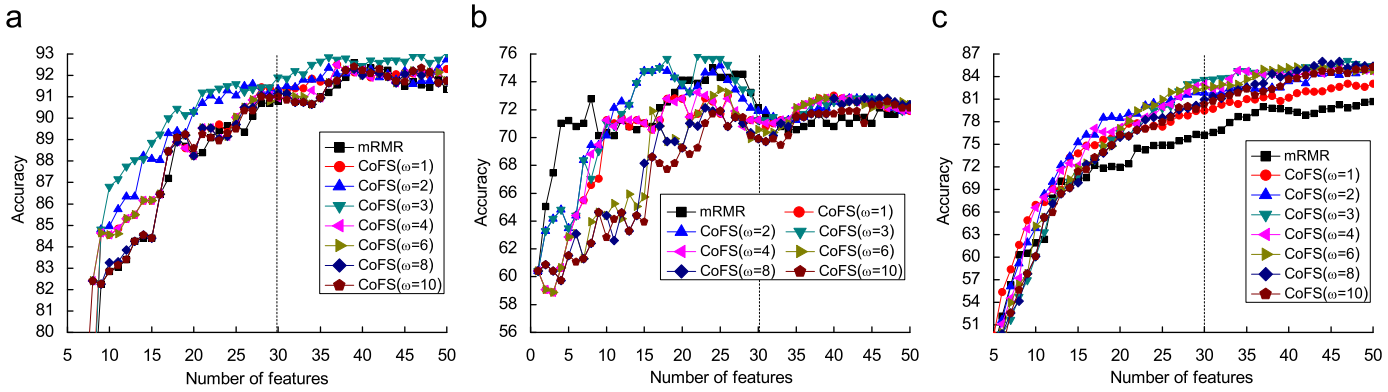

Fig. 3. Classification performance of SVM against different value of ω on three UCI datasets: (a) multi-feature pixel, (b) arrhythmia, (c) isolet.

Table 5
Running time (seconds) for each feature selection algorithm on UCI data sets.

Dataset	CoFS	mRMR
Synthetic	1.06	0.31
Molecular splice	2.59	0.92
Optical Recognition	6.42	2.63
Musk (Version 2)	29.68	14.10
Multi-feature factors	31.39	18.65
Multi-feature pixel	44.37	22.28
Arrhythmia	18.37	8.26
Isolet	124.02	57.95

method suggested an effective way to retain useful interdependent features and groups as many as possible.

From the previous empirical study, we can conclude that the proposed CoFS provides an efficient approach to evaluate the relevance of the features as a group with the target class. To illustrate this, we can see from Fig. 2 that CoFS did not select the first few features having the maximal relevance with the target, therefore, the classification accuracies are very low at the beginning. However, the CoFS method achieves more excellent performance than others after selecting a certain number of features. To take Synthetic dataset for an example, the accuracy of CoFS with the first feature is the lowest (only 26.25%) comparing with mRMR, ReliefF and IG (63.67%, 48.83%, and 63.67). Until the fourth feature is selected, the CoFS did not work well. However, when the fifth feature was included, the proposed CoFS method achieved the best accuracy and never had been surpassed from then on. It also can be seen from Fig. 2(b) that the top accuracy (92.21%) is achieved by CoFS using the first eight features.

In addition, it is noticeable that the proposed method performs different performances for different classifiers on the same dataset. Consequently, for different application fields, a suitable classifier is also necessary. And this issue is one of the most important challenges in the application of artificial intelligence.

6. Conclusions

Feature selection plays an important role in machine learning and data mining. In this paper, we firstly introduce a cooperative game theory based framework to evaluate the *power* of each feature, in order to overcome the disadvantage that traditional Information-theoretic based selectors ignore some features which as a group have strong discriminatory power but are weak as individuals. The second objective of this paper is to propose a general filter feature selection scheme based on the introduced framework to handle the feature selection problem. In this scheme, many Information-theoretic based feature selection methods, such as BIF, MIFS, SU and mRMR, can be employed. Experimental results on fifteen UCI datasets show that the proposed method works well and outperforms mRMR, Relief and IG at most cases. Its proven efficiency and effectiveness compared with other algorithms by four classifiers suggest that CoFS is practical for feature selection of high-dimensional data.

Acknowledgments

We would like to express our sincere appreciation to the anonymous reviewers for their insightful comments, which have greatly aided us in improving the quality of the paper.

This work is supported by the National Natural Science Foundation of China (No. 60973136, 61073164), Erasmus Mundus External Cooperation Window's Project (EMECW): Bridging the Gap (No. 155776-EM-1–2009-1-IT-ERAMUNDUS-ECW-L12), China-BC ICSD Grant (No. 2008DFA12140), Science Foundation for Youth of Jilin Province(201101033) and Science Foundation for Young Teachers of Jilin University (450060445169).

References

[1] Y. Kim, W.N. Street, F. Menczer, Feature selection in data mining, in: Wang John (Ed.), Data Mining, IGI Publishing, Hershey, 2003, pp. 80–105.
[2] Q. Zhu, L. Lin, M.L. Shyu, S.C. Chen, Feature selection using correlation and reliability based scoring metric for video semantic detection, in: Proceedings of the IEEE 4th International Conference on Semantic Computing, 2010, pp. 462–469.
[3] H. Ogura, H. Amano, M. Kondo, Comparison of metrics for feature selection in imbalanced text classification, Expert Systems with Applications 38 (5) (2011) 4978–4989.
[4] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.
[5] R. Cai, Z. Hao, X. Yang, W. Wen, An efficient gene selection algorithm based on mutual information, Neurocomputing 72 (4-6) (2009) 991–999.
[6] F. Amiri, M. Rezaei Yousefi, C. Lucas, A. Shakery,, N. Yazdani, Mutual information-based feature selection for intrusion detection systems, Journal of Network and Computer Applications 34 (4) (2011) 1184–1199.
[7] 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.
[8] S. Lin, Z. Lee, S. Chen, T. Tseng, Parameter determination of support vector machine and feature selection using simulated annealing approach, Applied Soft Computing 8 (4) (2008) 1505–1512.
[9] S.P. Moustakidis, J.B. Theocharis, Svm-fuzcoc: a novel svm-based feature selection method using a fuzzy complementary criterion, Pattern Recognition 43 (11) (2010) 3712–3729.
[10] H. Liu, L. Liu, H. Zhang, Boosting feature selection using information metric for classification, Neurocomputing 73 (1-3) (2009) 295–303.

[11] S. Kotsiantis, Feature selection for machine learning classification problems: a recent overview, Artificial Intelligence Review (2011) 1–20.
[12] J. Hua, W.D. Tembe,, E.R. Dougherty, Performance of feature-selection methods in the classification of high-dimension data, Pattern Recognition 42 (3) (2009) 409–424.
[13] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1-2) (1997) 273–324.
[14] M.M. Kabir, M.M. Islam, K. Murase, A new wrapper feature selection approach using neural network, Neurocomputing 73 (16-18) (2010) 3273–3283.
[15] I. Inza, P. Larrañaga, R. Etxeberria, B. Sierra, Feature subset selection by Bayesian network-based optimization, Artificial Intelligence 123 (1-2) (2000) 157–184.
[16] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning 46 (1) (2002) 389–422.
[17] L.C. Molina, L. Belanche,A. Nebot, Feature selection algorithms: a survey and experimental evaluation, in: Proceedings of the IEEE International Conference on Data Mining, 2002, pp. 306–313.
[18] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.
[19] L. Huan,, Y. Lei, Toward integrating feature selection algorithms for classification and clustering, IEEE Transactions on Knowledge and Data Engineering 17 (4) (2005) 491–502.
[20] H. Liu, H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Kluwer Academic Publishers, Norwell, Boston, 1998.
[21] K. Kira, L.A. Rendell, A practical approach to feature selection, in: Proceedings of the 9th International Workshop on Machine Learning, Morgan Kaufmann Publishers, 1992, pp. 249–256.
[22] I. Kononenko, Estimating attributes: analysis and extensions of relief, in: Proceedings of the 1994 European Conference on Machine Learning, 1994, pp. 171–182.
[23] S.W. Card, Information distance based fitness and diversity metrics, in: Proceedings of the 12th Annual Conference Companion on Genetic and Evolutionary Computation, Portland, OR, USA, 2010, pp. 1851–1854.
[24] K. Thangavel, A. Pethalakshmi, Dimensionality reduction based on rough set theory: a review, Applied Soft Computing 9 (1) (2009) 1–12.
[25] H. Liu, J. Li, L. Wong, A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns, Genome Informatics 13 (2002) 51–60.
[26] G. Brown A new perspective for information theoretic feature selection, in: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, 2009, pp. 49–56.
[27] M. Robnik-Sikonja, I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, Machine Learning 53 (1) (2003) 23–69.
[28] H. Liu, H. Motoda, L. Yu, A selective sampling approach to active feature selection, Artificial Intelligence 159 (1-2) (2004) 49–74.
[29] Y. Sun, Iterative relief for feature weighting: algorithms, theories, and applications, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (6) (2007) 1035–1051.
[30] Q. Hu, X. Che, L. Zhang, D. Yu, Feature evaluation and selection based on neighborhood soft margin, Neurocomputing 73 (10-12) (2010) 2114–2124.
[31] M. Dash, H. Liu, Consistency-based search in feature selection, Artificial Intelligence 151 (1-2) (2003) 155–176.
[32] J. Richard, Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches, IEEE Transactions on Knowledge and Data Engineering 16 (12) (2004) 1457–1471.
[33] Q. Hu, D. Yu, Z. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognition Letters 27 (5) (2006) 414–423.
[34] Y. Qian, J. Liang, W. Pedrycz, C. Dang, An efficient accelerator for attribute reduction from incomplete data in rough set framework, Pattern Recognition 44 (8) (2011) 1658–1670.
[35] D. Zhang, S. Chen, Z. Zhou, Constraint score: a new filter method for feature selection with pairwise constraints, Pattern Recognition 41 (5) (2008) 1440–1451.
[36] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks 5 (4) (1994) 537–550.
[37] K. Nojun, C. Chong-Ho, Improved mutual information feature selector for neural networks in supervised learning, in: Proceedings of the International Joint Conference on Neural Networks, 1999, pp. 1313–1318.
[38] N. Kwak, C. Choi, Input feature selection by mutual information based on parzen window, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (12) (2002) 1667–1671.
[39] J. Novovicova, P. Somol, M. Haindl, P. Pudil, Conditional mutual information based feature selection for classification task, in: Proceedings of the 12th Iberoamerican conference on Congress on Pattern Recognition, Chile, 2007, pp. 417–426.
[40] 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.
[41] L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, Journal of Machine Learning Research 5 (2004) 1205–1224.
[42] H. Liu, J. Sun, L. Liu, H. Zhang, Feature selection with dynamic mutual information, Pattern Recognition 42 (7) (2009) 1330–1339.
[43] Q. Hu, D. Yu, Z. Xie, J. Liu, Fuzzy probabilistic approximation spaces and their information measures, IEEE Transactions on Fuzzy Systems 14 (2) (2006) 191–201.
[44] H. Lee, C. Chen, J. Chen, Y. Jou, An efficient fuzzy classifier with feature selection based on fuzzy entropy, IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics 31 (3) (2001) 426–432.
[45] J. Shie, S. Chen, Feature subset selection based on fuzzy entropy measures for handling classification problems, Applied Intelligence 28 (1) (2008) 69–82.
[46] C. Lee, G.G. Lee, Information gain and divergence based feature selection for machine learning based text categorization, Information Processing & Management 42 (1) (2006) 155–165.
[47] T.M. Cover, J.A. Thomas, Elements of Information Theory, Second Edition., John Wiley & Sons, Inc., Hoboken, NJ, USA, 2005.
[48] S. Davies, S. Russell, *NP*-completeness of searches for smallest possible feature sets, AAAI Symposium on Intelligent Relevance, AAAI Press, New Orleans, 1994, pp. 41–43.
[49] J.W. Friedman, Game Theory with Applications to Economics, Second ed., Oxford University Press, New York, 1990.
[50] J.F. Banzhaf, Weighted Voting doesn't Work a Mathematical Analysis, 19, Rutgers Law Review, 1965 317–343.
[51] I. Guyon, A. Elisseeff, C. Aliferis, Causal feature selection, in: H. Liu, H. Motoda (Eds.), Computational Methods of Feature Selection, Chapman and Hall Press, 2007, pp. 63–82.
[52] P.D. Straffin, Game Theory and Strategy, Mathematical Association of America, Washington, DC, 1993, pp. 181–190.
[53] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms, 3rd edn., MIT Press, 2009.
[54] A. Frank, A. Asuncion, UCI Machine Learning Repository, Available from: 〈http://archive.ics.uci.edu/ml〉. Irvine, CA: University of California, School of Information and Computer Science, 2010.
[55] J. Grzymala-Busse, M. Hu, A comparison of several approaches to missing attribute values in data mining, Rough Sets and Current Trends in Computing 2005 (2001) 378–385.
[56] U.M. Fayyad, K.B. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, in: Proceedings of the 13th International Joint Conference on Artificial Intelligence, 1993, pp. 1022–1027.
[57] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z. Zhou, M. Steinbach, D. Hand,, D. Steinberg, Top 10 algorithms in data mining, Knowledge and Information Systems 14 (1) (2008) 1–37.
[58] C.E. Brodley, Recursive automatic bias selection for classifier construction, Machine Learning 20 (1) (1995) 63–94.

Xin Sun is currently a PhD. Candidate at the College of Computer Science and Technology, Jilin University, PR China. He received the MSc degrees from the College of Computer Science and Technology, Jilin University, PR China, in 2010. His research interests include pattern recognition, computer security and complex network.

Yanheng Liu received the PhD degree from the College of Computer Science and Technology, Jilin University, PR China, in 2003. He is currently a professor and supervisor of PhD candidates with Jilin University. He has wide research interests, mainly including artificial intelligence, pattern recognition, network security, complex networks, wireless sensor networks, mobile IP technology and QoS mechanism.

Jin Li is currently a PhD. Candidate at the school of philosophy and society, Jilin University, PR China. She received the MSc degrees from the institute of higher education, Jilin University, PR China, in 2010. Her research interests include theoretical psychology, group dynamics and game theory.

Jianqi Zhu received the PhD and MSc degree from the College of Computer Science and Technology, Jilin University, PR China in 2009 and 2004, respectively. Currently, he is a Lecturer in College of Computer Science and Technology, Jilin University, PR China. His research interests include network security, software watermarking and complex network.

Huiling Chen is currently a PhD Candidate at the College of Computer Science and Technology, Jilin University, PR China. He received his MS in Department of computer science and technology at Changchun University of technology, PR China, in 2008. His research interests include artificial intelligence, pattern recognition and machine learning.

Xuejie Liu received the PhD and MSc degree from the College of Computer Science and Technology, Jilin University, PR China in 2008 and 2004, respectively. Currently, She is a Lecturer in College of Computer Science and Technology, Jilin University, PR China. Her research interests include machine learning, network security and complex network.