Neurocomputing 97 (2012) 86–93

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Using cooperative game theory to optimize the feature selection problem

Xin Sun [a,b], Yanheng Liu [a,b,*], Jin Li [c], Jianqi Zhu [a,b], Xuejie Liu [a,b], Huiling Chen [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 28 October 2011
Received in revised form
10 February 2012
Accepted 2 May 2012
Communicated by X. Wu
Available online 29 May 2012

Keywords:
Machine learning
Feature selection
Shapley value
Filter method

ABSTRACT

Feature selection is an important preprocessing step in machine learning and pattern recognition. Recent years, various information theoretic based measurements have been proposed to remove redundant and irrelevant features from high-dimensional data set as many as possible. One of the main disadvantages of existing filter feature selection methods is that they often ignore some features which have strong discriminatory power as a group but are weak as individuals. In this work, we propose a new framework for feature evaluation and weighting to optimize the performance of feature selection. The framework first introduces a cooperative game theoretic method based on Shapley value to evaluate the weight of each feature according to its influence to the intricate and intrinsic interrelation among features, and then provides the weighted features to feature selection algorithm. We also present a flexible feature selection scheme to employ any information criterion to our framework. To verify the effectiveness of our method, experimental comparisons on a set of UCI data sets are carried out using two typical classifiers. The results show that the proposed method achieves promising improvement on feature selection and classification accuracy.

1. Introduction

Data mining is the process of analyzing data from different perspectives and extracting it into useful information. Along with the new emergences of computer applications, such as social networks clustering, gene expression array analysis and combinatorial chemistry, datasets are getting larger and larger. Nevertheless, most of the features in huge dataset are irrelevant or redundant, which lead traditional mining and learning algorithms to low efficiency and over-fitting. To mitigate this problem, one effective way is to reduce the dimensionality of feature space with feature selection technique [1]. Feature selection can bring lots of benefits to machine learning algorithms [2], such as reducing the measurement cost and storage requirements, coping with the degradation of the classification performance due to the finiteness of training sample sets, reducing training and utilization time, and facilitating data visualization and data understanding. It has attracted great attention and many selection algorithms have been developed during past years. Previous reviews of feature selection can be found in literatures [2–6]. Generally, all of these selection algorithms typically fall into three categories: embedded, wrapper and filter methods. Embedded and wrapper methods are specific to a given learning algorithm. For example, Guyon et al. [7] proposed a embedded method (SVM-RFE) utilizing support vector machine methods based on recursive feature elimination. And Cohen et al. [8,9] presented wrapper methods for feature selection based on the multi-perturbation Shapley analysis, in which the validation accuracy of a classifier was used to evaluate the contribution of each feature. One drawback of these methods 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. Filter methods are independent of learning algorithms and assess the relevance of features by looking only at the intrinsic properties of the data. In practice, filter methods have much lower computational complexity than others, meanwhile, they achieve comparable classification accuracy for most classifiers. So far, a modest number of efficient filter selection algorithms have been proposed in literatures. It is noteworthy that among various evaluation criteria, information theoretic based measurements achieve excellent performance and have drawn more and more attention (e.g., [10–16]). 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 some features having strong discriminatory power as a group but weak as individuals [2]. The main reason for this disadvantage is that information theoretic based measurements disregard the intrinsic structure [17] among features.

To untie this knot, we propose a new framework (shown in Fig. 1) to improve the performance of feature selection. First, the framework introduces a cooperative game theoretic method

[*] Corresponding author. Tel.: +86 431 85159419; fax: +86 431 85168337.
E-mail address: yhliu@jlu.edu.cn (Y. Liu).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.05.001

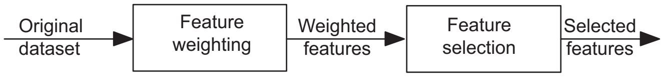

Fig. 1. A framework for feature selection.

to evaluate the weight of each feature according to its influence to the intricate and intrinsic interrelation among features. 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 feature can be weighted by averaging the contributions that it makes to each of the subset which it belongs to. After that the framework provides the weighted features to a feature selection algorithm. A scalable and flexible feature selection scheme is also presented to employ any information criterion to our framework.

This paper is organized as follows: Section 2 introduces some basic concepts of information theory and necessary background of cooperative game theory. Section 3 proposes a general framework for feature evaluation based on cooperative game theory, and presents a feature selection scheme to employ any existing information criterion. We focus on reducing the computational complexity of feature evaluation in Section 4. Section 5 gives experimental results on UCI data sets to evaluate the effectiveness of our method, and some discussions. Conclusions and future work are presented in Section 6.

2. Preliminaries

2.1. Entropy and mutual information

The fundamental quantities of information theory [18] – entropy and mutual information – provide intuitive tools to measure the uncertainty of random variables and the information shared by them.

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

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

Conditional mutual information (CMI) is also used as the reduction in the uncertainty of X due to knowledge of Y when Z is given.

2.2. Cooperative game theory

Coalitions: $N=\{1, 2, \ldots, n\}$ is the set of players. A coalition is a subset of players, N, which is able to make a binding agreement. Any subset of N, including N itself, can form a coalition [19].

A coalitional game is a pair (N, v), where N is a finite set of players, indexed by i; and $v:2^N\rightarrow R$ associates with each coalition $K\subseteq N$ a real-valued payoff $v(K)$ that the coalition's members can distribute among themselves, satisfying $v(\emptyset)=0$.

In game theory, the Shapley value was proposed by Shapley [20], which yields a unique outcome in coalitional games, to measure the powers of players in the game. In this paper, we use Shapley value to evaluate the weight of every feature.

3. Feature evaluation and selection

In natural large-scale data set, there are intrinsic correlations among features, such as causality, interdependence and unidirectional dependent. Among these intrinsic correlations, interdependent relationship commonly exists in real data set and is especially useful for classification tasks. However, most of the filter feature selection algorithms discard redundant features which are highly correlated with the selected ones. Consequently, interdependent features, weak as individuals but having strong discriminatory power as a group, are likely to be disregarded. Guyon et al. [5] also constructed an example by the famous XOR problem to illuminate that two variables which are useless by themselves can be useful together. To untie this knot, we preweight each feature according to its relationship (interdependence and redundancy) with others. And information theoretic measures, such as MI and CMI, are adopted in our relationship analysis.

3.1. Relevance, interdependence and redundancy analysis

3.1.1. Relevance

Relevance implies that the feature can contribute to the prediction accuracy [21]. For filter methods, correlation is widely used for relevance analysis [22]. In terms of information theory, the more relevant feature means it shares more information with target class. Mutual information (*MI*) is widely used to measure the relevance between two variables [10–13,15]. For example, the relevance of feature f for the target class is measured by *MI* as $I(f$, class), where $I(f$, class$)=0$ indicates that feature f is totally irrelevant with the target class.

3.1.2. Redundancy and interdependence

Two features are redundant to each other if their values are completely correlated [22]. As aforesaid discussion, most of the traditional feature selection methods disregard the intrinsic interdependent groups among features while eliminate redundant features. The main reason is that features which have been labeled "redundancy" may not be real redundancy. We tackle this problem by combining *MI* and *CMI* to distinguish the redundancy and interdependence of two features.

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. Meanwhile, a feature should also be considered redundant when it does not contribute whatsoever to others. Thus, feature f_i is said to be redundant with feature f_j if the following form is satisfied

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

Meanwhile, interdependence implies each member in the relationship cannot function or survive apart from one another, namely the impact of each feature on the classification performance cannot be ignored and replaced. Thus, suppose D is a set of features in an interdependent relationship, then the relevance between f_j and target class can be increased conditioned by $\forall f_i\in D$

($f_j \in D$ and $f_i \neq f_j$). Two features f_i and f_j are interdependent on each other if the following form is satisfied.

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

It is easy to see that the optimal feature subset is the one which all the features are relevant to the target class and interdependent on each other. Our contribution focus on evaluating the importance(or weight) of each feature, and retaining the useful interrelationships of features for feature selection.

To facilitate the following discussion, two criteria to measure the relevance of a set with the target class are presented. A relevance criterion, introduced by Peng et al. [13] as formula (6), is adopt to measure the relevance of set K on the target class.

$$D(K,class) = \frac{1}{|K|}\sum_{f_j \in K} I(f_j; class) \tag{6}$$

The change of the relevance of set K on the target class due to the knowledge of feature $f_i(f_i \notin K)$ is measured by the approximation

$$I(K; class; f_i) \approx \frac{1}{|K|}\sum_{f_j \in K} I(f_j; class|f_i) - I(f_j; class) \tag{7}$$

which is introduced by Meyer et al. [23].

3.2. Framework for feature evaluation via Shapley value

Shapley value measures the distribution of the power among the players in the voting game [19], which can be transformed into the arena of feature selection attempting to estimate the importance of each feature. In the context of feature selection, each coalition can be regarded as a candidate subset for the final selected optimal subset. The Shapley value provides a fair and efficient way to estimate the features' importance corresponding to the contribution of the features, while considering their possible intrinsic and intricate correlative interactions.

The original definition of Shapley value is described as follows [20]: The Shapley value itself is denoted $\phi(v)$, where $\phi(v) \in R^n$ and $\phi_i(v)$ is the Shapley value payoff to the ith player. The formula is

$$\phi_i(v) = \sum_{K \subset N} \Delta_i(K) x \frac{|K|!(n-|K|-1)!}{n!} \tag{8}$$

and

$$\Delta_i(K) = v(K \cup i) - v(K) \tag{9}$$

where n is the total number of players and the sum extends over all subsets K of N not containing player i.

For the sake of convenient, a interdependence index $\psi(i, j)$ is defined as

$$\psi(i,j) = \begin{cases} 1, & I(f_j; class|f_i) > I(f_j; class) \\ 0, & \text{else} \end{cases} \tag{10}$$

Based on the above definitions, we redefine the function $\Delta_i(K)$ associated with feature selection as

$$\Delta_i(K) = \begin{cases} 1, & I(K; class; f_i) \geq 0 \text{ and } \sum_{f_j \in K} \psi(i,j) \geq \frac{|K|}{2} \\ 0, & \text{else} \end{cases}, \tag{11}$$

which means the feature is crucial to win the coalition only if it both increases the relevance of the unitary subset K on the target class and is interdependent with at least half of the members.

More specifically, details of the feature evaluation method based on Shapley value are presented in Algorithm 1. The output of this evaluation framework is a vector $w(1{:}|F|)$ of which each element $w(i)$ represents the Shapley value of feature f_i.

Algorithm 1. Feature evaluation based on Shapley value.

```
Input: A training sample $O$ with feature space $F$ and the target $C$.
Output: $w(1{:}|F|)$: weight vector of $F$.
Initialize the weight $w(f)=0$ for each features;
For each feature $i \in F$ do
    Create all coalitions set$\{\pi_1, \ldots, \pi_t\}$ over $F \backslash i$;
    For each $\pi_j \in \{\pi_1, \ldots, \pi_t\}$ do
        Calculate the value of $\Delta_i(\pi_j)$;
    End
    Calculate the Shapley value $\phi_i(v)$;
    $w(i) = \phi_i(v)$
End
Normalized the vector $w(1{:}|F|)$.
```

It is noticed that Shapley value is only a metric estimating the importance of every feature based on the intrinsic correlative structures among features. In the following section, we try to optimize the feature selection problem with the weighted features, and present a general feature selection scheme of combining our evaluation framework with any information theoretic criteria.

3.3. A general feature selection optimization scheme

Feature selection algorithm uses the features as input and evaluates each of them according to its own criterion. On this basis, we utilize the weights of features to re-evaluate the features. For example, the priority of features with higher(or lower) weight will be raised(or reduced). Without loss of generality, to handle the feature selection problem as suggested by literature [2,16], a straight-forward optimization scheme is presented to employ any information criterion, such as BIF [1], SU [22] and mRMR [13], to our framework. The pseudo codes are outlined as Algorithm 2. We consider the feature selection procedure in a straight forward way. And the selection procedure will be terminated if the number of selected features is larger than the user-specified threshold σ.

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

```
Input: A sampling dataset $T=(F, O, C)$ and $w(1{:}|F|)$.
Output: Selected feature subset $S$.
Initialize parameters: $S=\emptyset$, $k=0$;
While $j < \sigma$ do
    For each feature $f_i \in F$ do
        Calculate its value of criterion $J(f_i)$;
        Calculate its victory criterion $V(f_i)=J(f_i) \times w(i)$;
    End
    Choose the feature $f_i$ with the largest $V(f_i)$;
    $F=F \backslash \{f_i\}$, $S=S \cup \{f_i\}$;
    $j=j+1$;
End
```

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 $J(S)$, such as SU and mRMR, is used to pick out the feature f_i by feature relevance analysis or relevance and redundancy analysis. We utilize the weight (normalized Shapley value) $w(i)$, 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. There exist a great many information criteria that can be employed as the criterion function in the proposed general feature selection scheme. As an illustration, mRMR and SU are employed as the criterion function in this work.

4. Computational complexity reduction

The problem of finding the Shapley value for the voting game is known to be #*P*-complete in the general case [24]. In particular, the calculation of the Shapley value for each feature requires summing over all possible subsets of features (step 3 of Algorithm 1), which is impractical in typical feature selection problems. In fact, the number of features falling in one interdependent group is much smaller than the total number of features in the real dataset. It is unnecessary to consider all coalitions for features, especially large coalitions. Nay more, the consideration of large coalitions may deteriorate the selector's performance, for the probability that the final subset contains whole features of large coalition is very low. Thus, we suggest a limit value ω being a bound on the coalition size. The formula (8) can be redefined as

$$\phi_i(v) = \sum_{K \subset \Pi_\omega} \Delta_i(K) \frac{|K|!(n-|K|-1)!}{n!} \quad (12)$$

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

Moreover, even the conditional feature z is theoretically total independent with the two variables x and y, conditional mutual information $I(x;y|z)$ is hardly equal zero in practical cases. This is because the input samples are inaccurate in most cases. Nevertheless, these conditional mutual information values are usually extremely small and should be regarded as noise. In experiments, these noises value can be replaced by zero. Therefore, we can further reduce time complexity by the knowledge of combinatorial mathematics and dynamic programming algorithm. In this work, a threshold value δ is defined to eliminate the noise values by

$$\delta = \alpha \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (I(f_j;class|f_i) - I(f_j;class)), \quad (13)$$

Where α is a regulated coefficient.

5. Experiments and results

To test the proposed method empirically, several experiments have been carried out to evaluate the performance of our method by employing two typical information criteria SU and mRMR and comparing with the original feature selection algorithm SU and mRMR. ReliefF is used as a baseline algorithm for comparison. ReliefF [25] is one of the most successful feature selectors and adopted Euclidean distance to assign a relevance weight to each feature.

In addition to feature selection algorithms, we employ two representative classifiers, i.e., Naive Bayes (NB) and SVM, which stand for quite different machine learning approaches [26]. For estimating the performance of classification algorithms, a three times of ten-fold cross-validation is used. The final results are their average values.

The experimental workbench is Weka (Waikato environment for knowledge analysis), which is a collection of machine learning algorithms for data mining tasks [27]. The parameters of classifiers for each experiment are set to default values of Weka. All experiments are conducted on a Core(TM)2 T6670, with a CPU clock rate of 2.20 GHz and 2 G main memory.

Table 1
Summary of datasets in our experiments.

No.	Dataset	Samples	Features	Classes
1	Synthetic	600	60	6
2	Optical recognition	5620	65	10
3	Musk (version 2)	6598	166	2
4	Multi-feature pixel	2000	240	10
5	Arrhythmia	452	279	16
6	Isolet	1559	618	26

5.1. Data sets and preprocessing

In this work, six real world datasets from UCI's machine learning data repository [28] are adopted in our simulation experiments, as shown in Table 1. These datasets contain various numbers of features and come from different domains, such as 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 [29]. 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 [30].

5.2. Feature selection and classification results

5.2.1. Classification accuracies

The effectiveness of a feature selection algorithm can be simply and directly measured by the classification performance on different datasets for classifiers. For the sake of convenience of experimental works, the features selected by different algorithms are arranged in a descending order according to their priorities. Then, we compare their classification accuracies against the number of features as shown in 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 performance of classifiers of the first k features. The parameters ω and α is set to 3 and 0.2, respectively. More detail experiments about the two parameters will be presented subsequently.

On the one hand, feature selection is to select smaller number of features from a huge feature space. However, in practice, too few features cannot function well. As an example, for KDD synthetic dataset, the accuracies are lower than 86% with the first 9 features. In fact, the accuracy can be 96.50% with 12 features by CGFS-mRMR and 91% with 15 features by mRMR which are also much smaller than the original number of features. On the other hand, selecting too many features is no meaning for feature reduction at all. Thus, the number of features selected for classification should be in an acceptable range. It can be seen from Fig. 2 that CGFS-mRMR and CGFS-SU outperform mRMR and SU selection methods within an acceptable number of features in most cases, which verifies that the proposed scheme CGFS can improve the performance of original information theoretic based feature selection algorithm.

To better illustrate that, we rank all the features and generate feature subsets by picking the top m features, where $m=1$,…, 30. One goal of feature selection is to select a smaller subset of the extracted features. Thus, the one that achieves the highest accuracy for the classifier is selected as the best subset. Statistics of the top classification accuracies for the two classifiers are listed in Tables 2 and 3, respectively. In each cell, the accuracy is followed by the number of selected features. The bold value means that it is the largest one among these feature selection

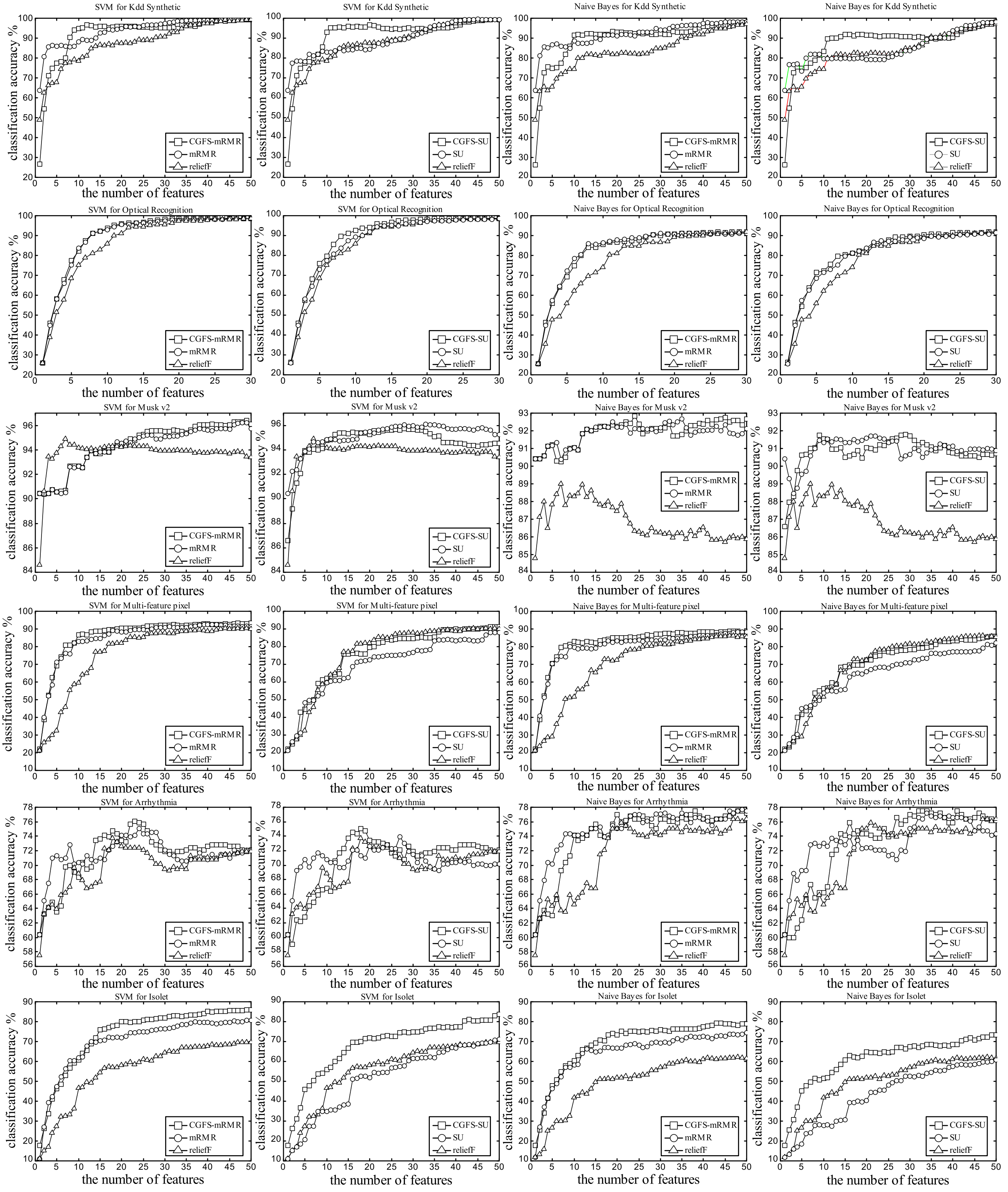

Fig. 2. Accuracies vs. different numbers of selected features on six UCI datasets by SVM and NB.

methods under the same classifier, and the average value of accuracies with the same selector is given in the row labeled as "Avg.".

As can be seen from Tables 2 and 3, the average value of classification accuracy (in the "Avg." row) for the six real datasets denotes that CGFS-mRMR and CGFS-SU perform better than the original mRMR and SU, while select less features. As an illustration, for CGFS-mRMR, it achieves the highest average accuracy (90.41%) by only using the fewest number of features. For the NB classifier, one may also observe that the CGFS scheme clearly surpasses others in most cases. To determine whether the experimental results are significant or not, paired t-tests between accuracies with the baseline algorithm ReliefF and with other selectors at a time have been carried out. 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

Table 2
The comparison of classification accuracies of SVM (the corresponding number of selected features in parentheses).

Dataset	CGFS-mRMR	mRMR	CGFS-SU	SU	ReliefF
KDD Synthetic	**96.50**%(12)**v**	94.67%(20)**v**	95.67%(13)**v**	84.83%(13)◦	87.67%(19)
Optical recognition	**98.72**%(19)**v**	97.65%(20)	98.24%(20)	96.76%(19)	97.35%(20)
Musk (Version 2)	95.56%(28)	95.21%(26)	95.92%(28)	**96.01**%(28)	94.92%(7)
Multi-feature pixel	**90.50**%(20)**v**	89.70%(25)	84.75%(24)	76.80%(30)◦	87.75%(30)
Arrhythmia	**76.10**%(23)**v**	75.00%(24)	75.33%(18)**v**	73.89%(27)	73.68%(18)
Isolet	**81.98**%(30)**v**	76.20%(30)**v**	74.98%(28)**v**	61.51%(30)	64.34%(30)
Avg.	**90.41**%(19)	88.07%(21)	87.48%(19)	81.63%(22)	84.28%(20)

Table 3
The comparison of classification accuracies of naive Bayes (the corresponding number of selected features in parentheses).

Dataset	CGFS-mRMR	mRMR	CGFS- SU	SU	ReliefF
Kdd Synthetic	**92.17**%(13)**v**	91.50%(17)**v**	91.67%(15)**v**	81.83%(7)	80.17%(15)
Optical recognition	90.80%(19)	**91.28**%(18)	90.25%(20)	89.16%(18)	89.27%(20)
Musk (Version 2)	**92.83**%(24)**v**	92.51%(19)**v**	91.54%(10)	91.72%(21)	89.00%(7)
Multi-feature pixel	**85.50**%(23)**v**	83.55%(25)	77.50%(24)	70.95%(30)◦	81.70%(30)
Arrhythmia	**78.23**%(20)	77.00%(24)	75.90%(16)	73.90%(26)◦	75.89%(21)
Isolet	**75.88**%(30)**v**	69.79%(30)	67.55%(30)**v**	52.73%(30)	56.90%(30)
Avg.	**85.90**%(20)	84.27%(21)	82.40%(18)	76.71%(21)	78.82%(20)

a paired *t*-test. Notation "**v**"(or "◦") represents that the value of current entry is significantly better (or worse) than the corresponding one in the "ReliefF" column in statistical *t*-test. As can be seen from Table 2, the original mRMR only achieves significantly better performance than ReliefF in two cases, meanwhile, SU even has two significantly worse cases. On the contrary, CGFS-mRMR achieves significantly better performance than ReliefF in five cases, and CGFS-SU has three significantly better cases with none worse case. Overall, our feature selection framework achieves promising improvement in the performance of classifiers.

5.2.2. *Parameters selection*

There are two parameters (ω and α) in our framework, which are introduced to reduce the computational complexity and improve the accuracy of feature evaluation.

The limit value ω 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. However, for datasets derived from real applications, the size is not large in most cases. 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 Synthetic dataset with $\alpha=0.2$. Fig. 3 shows the experimental results using $\omega=1$, $\omega=3$ and $\omega=5$. We can see that the classification accuracy reaches its highest when the limit value is set to 3 for Synthetic dataset, and achieves lower performance by rising ω to 5. Of course other values may perform better performance for a certain dataset, e.g., the classification performance gets slightly better with $\omega=2$ on optical recognition dataset. It is also can be seen that learning performance does not drop down dramatically as the value of ω increasing. The underlying reason perhaps is that the Shapley value (formula (6)) is a weighted sum of its marginal contributions to all the coalitions of which it is a member and weights a coalition inversely to the number of coalitions that exist of the given size [20]. Because a feature is crucial, on the whole, in smaller winning coalitions (because together with the most importance feature, fewer interdependence features are needed), the Shapley value will favor it more [19]. Therefore, we simply suggest the limit size a moderate value $\omega=3$ in all experiments in this paper.

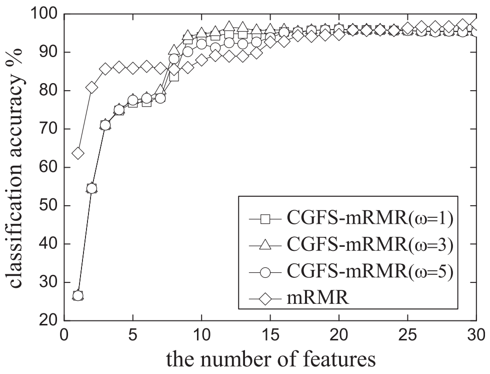

Fig. 3. Classification performance of SVM against different value of ω.

The parameter α is a regulated coefficient in formula (13). We conduct experiments with different value of the parameter α and a fixed value of parameter $\omega=3$. Results show that our method can achieve better performance with acceptable computational complexity when $\alpha\in[0.3, 0.5]$. Table 4 shows a part of our results of the highest accuracies and running time (seconds) of feature selection against different value of α.

5.3. *Discussions*

For datasets with a large number of features, there exist various intrinsic correlations among variables [14] including interdependence and causality. And the groups of features in interdependent relationship are useful for machine learning, such as co-regulated gene groups [31] in gene expression array dataset. In this work, we focus on exploring a new framework to retain the useful structure among features as many as possible. From the previous empirical study, we can conclude that the proposed feature selection framework provides an efficient approach to optimize the performance of feature selection. To illustrate this, we can see from Fig. 2 that the classification accuracies of CGFS are very low at the beginning in some cases (Synthetic, Musk V2, Arrhythmia). This is because CGFS does not select the first few features having the maximal relevance with the target. However,

Table 4
The comparison of highest accuracies and running time with different value of α.

Dataset	$\alpha=0$		$\alpha=0.2$		$\alpha=0.5$		$\alpha=0.8$		$\alpha=1$	
	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time
KDD Syntheticsynthetic	**96.50%**	196 s	**96.50%**	34 s	**96.50%**	7.8 s	c	2.1 s	91.42%	1.5 s
Optical recognition	97.52%	239 s	**98.36%**	102 s	97.68%	3.7 s	96.82%	1.3 s	93.95%	1.1 s
Musk (version 2)	92.00%	1286 s	95.56%	390 s	**96.13%**	220 s	95.10%	58 s	90.43%	39 s
Multi pixel	90.28%	> 2 day	90.50%	3182 s	**91.70%**	1638 s	89.64%	560 s	89.64%	240 s
Arrhythmia	74.90%	> 2 day	76.10%	4014 s	**78.00%**	1261 s	73.81%	52 s	65.30%	31 s
Isolet	80.35%	> 10 day	81.98%	51861 s	**83.65%**	2120 s	79.52%	380 s	71.66%	56 s

our method achieves more excellent performance than others after selecting a certain number of features. To take Synthetic dataset for an example, the accuracy of CGFS-mRMR with the first feature is the lowest (only 26.5%) comparing with mRMR and ReliefF (63.67% and 48.83%). Until the fourth feature is selected, the CGFS-mRMR did not work well. However, when the seventh feature was included, our method achieved the best accuracy and never had been surpassed from then on. It also can be seen that the top accuracy (96.50%) is achieved by CGFS-mRMR using the first twelve features.

Not all datasets from various domains are full of prominent interdependent groups. From Fig. 2 we can see that our method only achieve slight improvement on optical recognition dataset. Perhaps the main reason centers on the weak dependence among features. However, along with emergence of social networks and bio-informatics, datasets are getting more complex and complicated. It is helpful to discover and select the most interdependent groups for machine learning. As we all known, it is a hard problem to discover the association relationship among features exactly. We cannot guarantee our approach retain all useful interdependent groups or the whole interdependent group, however, we suggest a effective way to retain useful interdependent features and groups as many as possible.

In addition, it is noticeable that the proposed method performs different performances for the two classifiers on the same dataset. For different application fields, exploring an optimal combination of feature selection algorithm and classifiers is critical. And this issue is one of the most important challenges in the application of artificial intelligence.

6. Conclusions and future work

Interdependent groups of features commonly exist in the real dataset and traditional feature selection algorithms always destroy these useful intrinsic groups. To overcome this disadvantage, we explored a general framework for feature evaluation and weighting to optimize the performance of feature selection. The framework first introduces a cooperative game theoretic method to evaluate the weight of each feature according to its influence to the intricate and intrinsic interrelation among features, e.g., the features which fall in an interdependent relationship will be assign higher weight. Then the weighted features are provided to the feature selection algorithm. Finally, a flexible feature selection scheme is proposed in order to employ any successful information criterion (such as BIF, SU and mRMR) to our framework. Experimental results on six UCI datasets show that the proposed method improves the performance of representative feature selection algorithms. Its proven efficiency and effectiveness compared with other algorithms by two classic classifiers suggest that the proposed framework is practical for feature selection of complex dataset.

Computational complexity is very important for feature selection of high-dimensional data. One of our future research directions is to further reduce the computational complexity by adopting approximate Shapley value estimate technique.

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): One-More-Step (No. 2011-2581 001-001 EMA2), Science Foundation for Youth of Jilin Province (No. 201101033) and Graduate Innovation Fund of Jilin University.

References

[1] A.K. Jain, R.P.W. Duin, J.C. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 4–37.
[2] X. Sun, Y. Liu, J. Li, J. Zhu, H. Chen, X. Liu, Feature evaluation and selection with cooperative game theory, Pattern Recognition 45 (2012) 2992–3002.
[3] H. Liu, L. Yu, Toward integrating feature selection algorithms for classification and clustering, IEEE Trans. Knowl. Data Eng. 17 (2005) 491–502.
[4] L.C. Molina, L. Belanche, A. Nebot, Feature selection algorithms: a survey and experimental evaluation, in: Proceedings of IEEE International Conference on Data Mining, IEEE Computer Society, Maebashi City, Japan, 2002, pp. 306–313.
[5] I. Guyon, E. Andr, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[6] Y. Kim, W.N. Street, F. Menczer, Feature selection in data mining, in: J. Wang (Ed.), Data Mining, IGI Publishing, 2003, pp. 80–105.
[7] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.
[8] S. Cohen, E. Ruppin, G. Dror, Feature selection based on the Shapley value, in: Proceedings of the 19th international joint conference on Artificial intelligence, Morgan Kaufmann Publishers, 2005, pp. 665-670.
[9] S. Cohen, G. Dror, E. Ruppin, Feature selection via coalitional game theory, Neural. Comput. 19 (2007) 1939–1961.
[10] N. Kwak, C.-H. Choi, Input feature selection by mutual information based on Parzen window, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 1667–1671.
[11] F. Fleuret, Fast binary feature selection with conditional mutual information, J. Mach. Learn. Res. 5 (2004) 1531–1555.
[12] V. Gómez-Verdejo, M. Verleysen, J. Fleury, Information-theoretic feature selection for functional data classification, Neurocomputing 72 (2009) 3580–3589.
[13] 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 (2005) 1226–1238.
[14] K.E. Hild, D. Erdogmus, K. Torkkola, J.C. Principe, Feature extraction using information-theoretic learning, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1385–1392.
[15] R. Cai, Z. Hao, X. Yang, W. Wen, An efficient gene selection algorithm based on mutual information, Neurocomputing 72 (2009) 991–999.
[16] D. Huang, T.W.S. Chow, Effective feature selection scheme using mutual information, Neurocomputing 63 (2005) 325–343.
[17] I. Guyon, A. Elisseeff, C. Aliferis, Causal feature selection, in: H.M.H. Liu (Ed.), Computational Methods of Feature Selection, Chapman and Hall Press, 2007, pp. 63–82.
[18] T.M. Cover, J.A. Thomas, J. Wiley, Elements of Information Theory, second ed., John Wiley & Sons, Inc., Hoboken, NJ, USA, 2005.

[19] J.W. Friedman, Game Theory with Applications to Economics, Oxford University Press, New York, 1990.
[20] L.S. Shapley, A value for n-person games, in: A.W.T.H.W. Kuhn (Ed.), In Contributions to the Theory of Games, vol. II, Princeton University Press, 1953, pp. 307–317.
[21] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.
[22] L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learn. Res. 5 (2004) 1205–1224.
[23] P.E. Meyer, C. Schretter, G. Bontempi, Information-theoretic feature selection in microarray data using variable complementarity, IEE, J. Sel. Top. Signal Process. 2 (2008) 261–274.
[24] X. Deng, C.H. Papadimitriou, On the complexity of cooperative solution concepts, Math. Oper. Res. 19 (1994) 257–266.
[25] I. Kononenko, Estimating attributes: analysis and extensions of ReliefF, Mach. Learn. 784 (1994) 171–182.
[26] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, D. Steinberg, Top 10 algorithms in data mining, Knowl. Inf. Syst. 14 (2008) 1–37.
[27] I.H. Witten, E. Frank, Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, Amsterdam, 2005.
[28] A. Frank, A. Asuncion, UCI Machine Learning Repository, Available 〈http://archive.ics.uci.edu/ml〉. Irvine, CA: University of California, School of Information and Computer Science, 2010.
[29] J. Grzymala-Busse, M. Hu, A comparison of several approaches to missing attribute values in data mining, Rough Sets Curr. Trends Comput. 2005 (2001) 378–385.
[30] U.M. Fayyad, K.B. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, in: Proceedings of 13th International Joint Conference on Artificial Intelligence, (Morgan Kaufmann, Chambéry, France, 1993), pp. 1022–1027.
[31] A. Gyenesei, U. Wagner, S. Barkow-Oesterreicher, E. Stolte, R. Schlapbach, Mining co-regulated gene profiles for the detection of functional associations in gene expression data, Bioinformatics 23 (2007) 1927–1935.

Xin Sun is currently a Ph.D. Candidate at the College of Computer Science and Technology, Jilin University, P.R. China. He received the M.Sc. degrees from the College of Computer Science and Technology, Jilin University, P.R. China, in 2010. His research interests include pattern recognition, computer security and complex network.

Yanheng Liu received the Ph.D. degree from the College of Computer Science and Technology, Jilin University, P.R. 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 Ph.D. Candidate at the school of philosophy and society, Jilin University, P.R. China. She received the M.Sc. degrees from the institute of higher education, Jilin University, P.R. China, in 2010. Her research interests include theoretical psychology, group dynamics and game theory.

Jianqi Zhu received the Ph.D. and M.Sc. degree from the College of Computer Science and Technology, Jilin University, P.R. China in 2009 and 2004, respectively. Currently, he is a Lecturer in College of Computer Science and Technology, Jilin University, P.R. China. His research interests include network security, software watermarking and complex network.

Xuejie Liu received the Ph.D. and M.Sc. degree from the College of Computer Science and Technology, Jilin University, P.R. China in 2008 and 2004, respectively. Currently, She is a Lecturer in College of Computer Science and Technology, Jilin University, P.R. China. Her research interests include machine learning, network security and complex network.

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