Neurocomputing 193 (2016) 33–41

Contents lists available at ScienceDirect

Neurocomputing

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

An efficient feature selection algorithm for hybrid data

Feng Wang [a,b], Jiye Liang [a,b,*]

[a] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, China
[b] School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China

ARTICLE INFO

Article history:
Received 10 July 2015
Received in revised form
5 January 2016
Accepted 15 January 2016
Communicated by Swagatam Das
Available online 20 February 2016

Keywords:
Feature selection
Hybrid data
Rough set theory
Large-scale data sets

ABSTRACT

Feature selection for large-scale data sets has been conceived as a very important data preprocessing step in the area of machine learning. Data sets in real databases usually take on hybrid forms, i.e., the coexistence of categorical and numerical data. In this paper, based on the idea of decomposition and fusion, an efficient feature selection approach for large-scale hybrid data sets is studied. According to this approach, one can get an effective feature subset in a much shorter time. By employing two common classifiers as the evaluation function, experiments have been carried out on twelve UCI data sets. The experimental results show that the proposed approach is effective and efficient.

1. Introduction

With the rapid development of information technologies including internet and databases, a very large number of data are acquired in many areas or industries, such as text and bioinformatics data. Both the size and the dimension of these data increase at an unprecedented rate, which has resulted in large-scale data with high dimension. Feature selection is an important technique used in dimensional reduction [3,4,6,18,19,22]. It aims to improve the accuracy and performance of classifiers through removing redundant features and selecting informative features from the data. Feature selection has been successfully used in many areas and has attracted much attention in recent years [13,14,32,38]. The rapid growth of data brings new challenges for traditional feature selection, and exploring efficient feature selection approaches has quickly become a key issue in machine learning [20,39,43,44,47,54].

In the process of feature selection, feature evaluation criteria are used to evaluate the quality of the candidate subsets. For a feature subset, different evaluation criteria may give different results. Roughly speaking, there are five kinds of evaluation criteria [8,23]. They are distance measures, information measures, dependency measures, consistency measures, and classification error rate measures. The first four evaluation criteria are used to evaluate feature subsets according to inherent characteristics of the data. The last one relies on a classification algorithm to evaluate and select useful features [1,25]. Obviously, compared with the first four evaluation criteria, using the last evaluation criteria can usually improve classification performance, but is also time-consuming. Because of different evaluation criteria being used in feature selection algorithms, existing feature selection algorithms are divided into three categories: the filter model, the wrapper model, and the hybrid model [3,18,25]. The "filter" algorithm model relies on the aforementioned first four evaluation criteria to select features. The "wrapper" algorithm model uses classification algorithms to evaluate candidate features. The hybrid model combines the advantages of "filter" and "wrapper" models by employing different evaluation criteria in a feature selection algorithm. With the development of feature selection and its deep research, the algorithms which can be used to deal with labeled data are called supervised feature selection algorithms [3]. The algorithms used to deal with unlabeled data are called unsupervised feature selection algorithms [5]. In addition, with the rise of big data, semi-supervised feature selection algorithms are gradually introduced to handle the small-labeled-sample problem in which unlabeled data is much more than labeled data [2]. For given data sets, feature types include numeric and nominal. To deal with nominal data, Wang [48] introduced a feature selection algorithm based on mutual information. Hu and Cercone [7] proposed another feature selection algorithm in which dependency is employed to measure the relevance between feature and class information. On this basis, consistency measure is introduced to evaluate and select features [4]. To deal with numeric data, discretization is a kind of common approach [8,24]. The domains of features are firstly segmented into some intervals by using discretization. Then, one can use above algorithms to select useful features. Hence, using different discretization approaches may lead to

* Corresponding author at: Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, China.
Tel./fax: +86 0351 7018176.
E-mail addresses: sxuwangfeng@126.com (F. Wang), ljy@sxu.edu.cn (J. Liang).

http://dx.doi.org/10.1016/j.neucom.2016.01.056
0925-2312/

different feature selection results. To overcome this limitation, distance measure is introduced to characterize the class separability.

In conclusion, many efficient feature selection approaches have been developed, and are very helpful for selecting features and saving computational time. However, most of them focus on dealing with data with a single type of features, i.e., nominal or numeric features [28,29,37,45,46,51]. In real databases, data sets usually take on hybrid forms. Nominal and numeric features coexist in real-world applications such as analysis of medical case, financial data and biological data. Therefore, research on feature selection for hybrid data has also gradually become a significant issue in data mining [36,42,48]. Several techniques used to select effective features from hybrid data sets were developed by some researchers [8–11,21,41,55], which are introduced as follows. By defining correlation between numeric features and correlation between nominal features respectively, Hall [8] and Yu and Liu [55] proposed the correlation-based feature selection algorithms for hybrid data. Tang and Mao [41] introduced another novel search algorithm. In this algorithm, based on nominal features, a hybrid data set is firstly decomposed into a series of feature subspaces. The class separability is measured according to numeric features in each subspace, and then combining all measures to produce an overall evaluation. Tang's algorithm give priority to classification ability of nominal features obviously. To overcome this limitation, based on the introduction of Parzen window, Kwak and Choi [21] proposed a new feature selection algorithm based on mutual information. In addition, as the fast development of rough set theory, some researchers began to study rough feature selection algorithms for hybrid data. Hu et al. [9] introduced a neighborhood rough set model, and gave a forward search algorithm which can be used to select feature subset from hybrid data. On the basis of fuzzy set theory, Hu et al. [10,11] constructed fuzzy equivalence relation matrix, and proposed an entropy-based feature selection algorithm for hybrid data. According to above fuzzy equivalence relation, Wei et al. [49] introduced another information entropy, and proposed an accelerated feature selection algorithm which can find a feature subset from hybrid data efficiently. Because the research is still in its infancy, some existing effective feature selection algorithms are usually low in computational efficiency, especially when dealing with large-scale hybrid data sets.

In this paper, based on the idea of decomposition and fusion, an efficient feature selection algorithm for large-scale hybrid data sets is studied. For a large-scale data set, collecting a part of records (or instances) from it can form a small data set, i.e., a subset. According to the idea of sample estimation, one can estimate on subsets the feature selection result of the original large-scale one. On this basis, a technique of decomposition and fusion is introduced to construct an efficient feature selection algorithm in this paper. Here, "decomposition" here means decomposing a large data set into a family of subsets which have the similar distribution with the large one (see Fig. 1). Feature selection results of multiple subsets can be considered as multiple estimates. Then, "fusion" means fusing all the estimates obtained from subsets together and generating a final feature subset of the large data set. Obviously, the total time spent on selecting features for subsets is much less than that for the original large-scale one, then this algorithm yields in a much less amount of time a feature subset. By employing two common classifiers (Naive Bayes classifier and decision tree classifier) as the evaluation function, experiments have been carried out on seven UCI data sets. The experimental results show that the proposed approach is effective and efficient.

The rest of this paper is organized as follows: some preliminaries are briefly reviewed in Section 2. In Section 3, by introducing an approach for decomposing a given large-scale hybrid data set and fusing all results, an efficient feature selection algorithm for large-scale hybrid data sets is proposed. In Section 4, twelve UCI large-scale data sets are employed to illustrate the effectiveness of the proposed algorithm. Section 5 concludes the paper with some discussions.

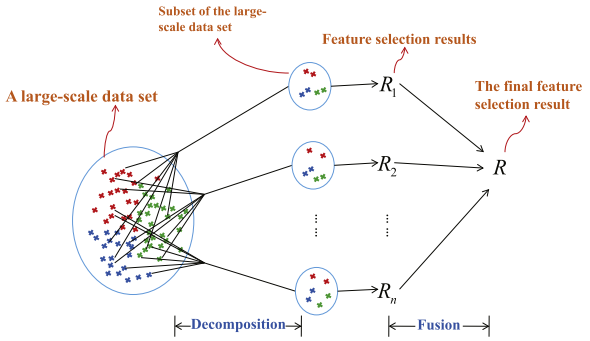

Fig. 1. Decomposition and fusion.

2. Preliminary knowledge

Pawlak defined the concept of rough sets in 1982 [33–35], which has been conceived as a powerful tool to deal with various types of data [26,27,30,31,40,50,52,53]. In rough set theory, a finite and nonempty set of instances U is called universe, and it is characterized with a set of attributes (features) A. Data table $S=(U,A)$ is called an information system in the rough set model. An attribute subset can induce an equivalence relation R on U and generates a family of equivalence classes. Instances with same attribute values are grouped into one equivalence class. A labeled data set is called a decision table $S=(U,C\cup D)$, where D is called decision attribute which denotes the column of class label and the rest of the features are called conditional attribute set C. The classic rough set model is only suitable for data with nominal values. To deal with numeric data in real world applications, researchers introduced fuzzy set into rough set and constructed fuzzy rough set model [9,12,13]. A numeric attribute induces a fuzzy equivalence relation instead of crisp equivalence relation. The fuzzy-rough set model is fitted for the case where both the relation and the object subset to be approximated are fuzzy. Some primary definitions of fuzzy rough set model are introduced as follows.

Let X be a non-empty finite set, and R is a binary relation defined on X, denoted by a relation matrix $M(R)$:

$$M(R)=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & & \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}, \tag{1}$$

where $r_{ij}\in[0,1]$ is the relation value of x_i and $x_j(x_i,x_j\in X)$. Furthermore, R is a fuzzy equivalence relation if R satisfies the following conditions:

(1) Reflectivity: $R(x,x)=1,\ \forall x\in X$;
(2) Symmetry: $R(x,y)=R(y,x),\ \forall x,y\in X$;
(3) Transitivity: $R(x,z)\geq \min_y\{R(x,y),R(y,z)\},\ \forall x,y,z\in U$.

The fuzzy equivalence class $[x_i]_R$ of x_i induced by the relation R is defined as

$$[x_i]_R=\frac{r_{i1}}{x_1}+\frac{r_{i2}}{x_2}+\cdots+\frac{r_{in}}{x_n},$$

where "+" means the union. The cardinality $[x_i]_R$ is defined as $|[x_i]_R|=\sum_{j=1}^{n} r_{ij}$.

Definition 1. Let $S=(U,A)$ be a fuzzy information system and $B,P\subseteq A$. Moreover, $[x_i]_B$ and $[x_i]_P$ are fuzzy equivalence classes

containing x_i generated by B and P, respectively. Then, the conditional entropy of B relative to P is defined as

$$H(P|B) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_P \cap [x_i]_B|}{|[x_i]_B|}.$$

Definition 2. Let $S=(U, C\cup D)$ be a fuzzy decision table and $B\subseteq C$. Then, the significance of $a\in C-B$ relative to D is defined as

$$SIG(a,B,D) = H(D|B) - H(D|B\cup\{a\}).$$

For a hybrid data table $S=(U,A)$, $A=A_1\cup A_2$. The relation matrix $M(R)=(r_{ij})_{n\times n}$ $(n=|U|)$ is computed as: 2 mm $\forall a\in A_1$ (numeric feature),

$$r_{ij}^a = \begin{cases} 1-4|a(x_i)-a(x_j)|, & |a(x_i)-a(x_j)|\le 0.25; \\ 0, & \text{otherwise.}\end{cases}$$

$\forall a\in A_2$ (nominal feature),

$$r_{ij}^a = \begin{cases} 1, & a(x_i)=a(x_j); \\ 0, & \text{otherwise.}\end{cases}$$

Then, $r_{ij}^A = \bigwedge_{a\in A} r_{ij}^a$.

Definition 3. Let $S=(U, C\cup D)$ be a fuzzy decision table and $B\subseteq C$. Then, B is a reduct of C relative to D if B satisfies:

(1) $H(D|B)=H(D|C)$;
(2) $\forall a\in B: H(D|B-a) > H(D|B)$.

3. Efficient feature selection for large-scale hybrid data sets

As mentioned above, rapid growth of data results in storage of a very large number of data, thus demanding more efficient approaches in data mining. In this section, three key problems, i.e., decomposition, selecting informative features from subsets and fusion, are introduced specifically. Then, an efficient feature selection algorithm for large-scale hybrid data is proposed. Because of that a data set is called a table in rough set theory, a subset of a given large-scale data set is called a "sub-table" in the following sections. In addition, it should be pointed out that we mainly study supervised feature selection in this paper, thus discussing the three problems based on labeled data sets in following subsections.

3.1. Decomposition of large-scale hybrid data sets

In the process of sample estimation, sample size needs to be determined firstly. In statistics, a familiar approach (see Definition 4) is using variance to determine the sample size for the given data set.

Definition 4. Let S be a data table (the original large-scale data table) and let the size of S be denoted by N. Then, the sample size M' is defined as [15,17]:

$$M' = \frac{Z^2\times s}{E^2}. \tag{2}$$

Where s means variance on S, Z means Z-statistic under confidence intervals, and E means margin error which can be adjusted as requested. This approach is very common in statistics, which has been widely used to estimate sample size in many instances such as estimating the annual salary, average consumption and average deposit.

To deal with hybrid data sets, here we propose an expansion of the variance s. Let $S=(U,A)$ be a hybrid data table, $a\in A$, $x_i\in U$ and $x_i=(a_1(x_i), a_2(x_i), \ldots, a_{|A|}(x_i))$. Supposed that $A=A_1\cup A_2$, A_1 includes numeric features and A_2 includes nominal features, then variance s_h on U is defined as:

$$s_h = \frac{|A_1|s_1 + |A_2|s_2}{|A|}, \tag{3}$$

where s_1 is the variance on numeric data [15], and s_2 is the variance on nominal data [16]. In addition, s_1 and s_2 are respectively defined as

$$s_1 = \frac{1}{|U|-1}\sum_{i=1}^{|U|}\left(\sum_{j=1}^{|A|} a_j(x_i) - a_j(\overline{x})\right)^2$$

and

$$s_2 = \frac{1}{k}\sum_{i=1}^{k}(|X_i| - \overline{X})^2,$$

where $a_j(\overline{x}) = \frac{\sum_{i=1}^{|U|} a_j(x_i)}{|U|}$, $U/A_2=\{X_1, X_2, \ldots, X_k\}$ and $\overline{X} = \frac{\sum_{i=1}^{k}|X_i|}{k}$.

Based on above introduction, Definition 5 shows the definition of sample size in a hybrid data set.

Definition 5. Let S be a hybrid data table, the sample size M' is defined as:

$$M' = \frac{Z^2\times s_h}{E^2}. \tag{4}$$

In addition, if sample size M' is larger than 5% of the overall size N, the sample size M' needs to be adjusted. In [17], the adjusted formula is defined as follows:

$$M_1 = \frac{M'N}{M'+N}. \tag{5}$$

It can be seen from formula (2) that computing s_h of a large-scale hybrid data table is obviously time-consuming. Therefore, two theorems are introduced as follows, which will be used in our further development.

Theorem 1. *Let $S_1=(U_1,A)$ and $S_2=(U_2,A)$ be two tables with numeric data. The average and variance on U_1 and U_2 are $\overline{\mathbf{x}}_1$, s_{11}, $\overline{\mathbf{x}}_2$ and s_{12}. Then, the average $\overline{\mathbf{x}}$ and variance s_1 on $U_1\cup U_2$ are defined as*

$$\overline{\mathbf{x}} = \frac{n\overline{\mathbf{x}}_1 + m\overline{\mathbf{x}}_2}{n+m}, \quad s_1 = \frac{(n-1)s_{11}+(m-1)s_{12}}{n+m-1} + \Delta_1,$$

where $\Delta_1 = \frac{nm(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2)^2}{(n+m)(n+m-1)}$, $n=|U_1|$, and $m=|U_2|$.

Proof. The proof can be found in [15].□

Theorem 2. *Let $S_1=(U_1,A)$ and $S_2=(U_2,A)$ be two tables with nominal data, $U_1/A=\{X_{11}, X_{12}, \ldots, X_{1k_1}\}$, and $U_2/A=\{X_{21}, X_{22}, \ldots, X_{2k_2}\}$. The variance on U_1 and U_2 are s_{21} and s_{22}. Suppose that $(U_1\cup U_2)/A=\{X_1, X_2, \ldots, X_l, X_{1l+1}, X_{1l+2}, \ldots, X_{1k_1}, X_{2l+1}, X_{2l+2}, \ldots, X_{2k_2}\}$, where $X_i = X_{1i}\cup X_{2i} (i=1,2,\ldots,l)$, then the variance s_2 on $U_1\cup U_2$ is defined as*

$$s_2 = \frac{k_1 s_{21} + k_2 s_{22}}{k_1+k_2-l} + \Delta_2,$$

where $\Delta_2 = \frac{1}{k_1+k_2-l}\left(\frac{n^2}{k_1}+\frac{m^2}{k_2}-\frac{(n+m)^2}{k_1+k_2-l}+2\sum_{i=1}^{l}|X_{1i}|\cdot|X_{2i}|\right)$, $n=|U_1|$, and $m=|U_2|$.

Proof. Let $\overline{X}_1$, $\overline{X}_2$ and $\overline{X}$ be average on U_1, U_2 and $U_1\cup U_2$. Then, one can get $\overline{X}_1 = \frac{n}{k_1}$, $\overline{X}_2 = \frac{m}{k_2}$ and $\overline{X} = \frac{n+m}{k_1+k_2-l}$. Let $k=k_1+k_2-l$,

variance s_2 can be defined as:

$$
\begin{aligned}
s_2 &= \frac{1}{k}\left(\sum_{i=1}^{l}(|X_i|-\overline{X})^2+\sum_{i=l+1}^{k_1}(|X_{1i}|-\overline{X})^2+\sum_{i=l+1}^{k_2}(|X_{2i}|-\overline{X})^2\right)\\
&= \frac{1}{k}\left(\sum_{i=1}^{l}(|X_{1i}|+|X_{2i}|-\overline{X})^2+\sum_{i=l+1}^{k_1}(|X_{1i}|-\overline{X})^2+\sum_{i=l+1}^{k_2}(|X_{2i}|-\overline{X})^2\right)\\
&= \frac{1}{k}\left(\sum_{i=1}^{l}(|X_{1i}|-\overline{X})^2+\sum_{i=1}^{l}(|X_{2i}|-\overline{X})^2+\sum_{i=1}^{l}(2|X_{1i}\|X_{2i}|-\overline{X}^2)\right.\\
&\quad \left.+\sum_{i=l+1}^{k_1}(|X_{1i}|-\overline{X})^2+\sum_{i=l+1}^{k_2}(|X_{2i}|-\overline{X})^2\right)\\
&= \frac{1}{k}\left(\sum_{i=1}^{k_1}(|X_{1i}|-\overline{X})^2+\sum_{i=1}^{k_2}(|X_{2i}|-\overline{X})^2+\sum_{i=1}^{l}(2|X_{1i}\|X_{2i}|-\overline{X}^2)\right)\\
&= \frac{1}{k}\left(\sum_{i=1}^{k_1}(|X_{1i}|-\overline{X}_1+\overline{X}_1-\overline{X})^2+\sum_{i=1}^{k_2}(|X_{2i}|-\overline{X}_2+\overline{X}_2-\overline{X})^2\right.\\
&\quad \left.+\sum_{i=1}^{l}(2|X_{1i}\|X_{2i}|-\overline{X}^2)\right)\\
&= \frac{1}{k}\left(\sum_{i=1}^{k_1}(|X_{1i}|-\overline{X}_1)^2+k_1(\overline{X}_1-\overline{X})^2+\sum_{i=1}^{k_2}(|X_{2i}|-\overline{X}_2)^2+k_1(\overline{X}_2-\overline{X})^2\right.\\
&\quad \left.+\sum_{i=1}^{l}(2|X_{1i}\|X_{2i}|-\overline{X}^2)\right)\\
&= \frac{1}{k}\left(k_1s_{21}+k_2s_{22}+k_1\overline{X}_1^2-2k_1\overline{X}_1\overline{X}+k_1\overline{X}^2\right.\\
&\quad \left.+k_2\overline{X}_2^2-2k_2\overline{X}_2\overline{X}+k_2\overline{X}^2-l\overline{X}^2+\sum_{i=1}^{l}2|X_{1i}\|X_{2i}|\right)\\
&= \frac{1}{k}\left(k_1s_{21}+k_2s_{22}+k_1\overline{X}_1^2+k_2\overline{X}_2^2-2k_1\overline{X}_1\overline{X}-2k_2\overline{X}_2\overline{X}+\overline{X}^2(k_1+k_2-l)\right.\\
&\quad \left.+\sum_{i=1}^{l}2|X_{1i}\|X_{2i}|\right)\\
&= \frac{1}{k}\left(k_1s_{21}+k_2s_{22}+k_1\overline{X}_1^2+k_2\overline{X}_2^2-2k_1\overline{X}_1\overline{X}\right.\\
&\quad \left.-2k_2\overline{X}_2\overline{E}+\overline{X}(n+m)+\sum_{i=1}^{l}2|X_{1i}\|X_{2i}|\right)\\
&= \frac{1}{k}\left(k_1s_{21}+k_2s_{22}+k_1\overline{X}_1^2+k_2\overline{X}_2^2-\overline{X}(2k_1\overline{X}_1+2k_2\overline{X}_2-n-m)\right.\\
&\quad \left.+\sum_{i=1}^{l}2|X_{1i}\|X_{2i}|\right)\\
&= \frac{1}{k}\left(k_1s_{21}+k_2s_{22}+k_1\overline{X}_1^2+k_2\overline{X}_2^2-\overline{X}(n+m)+\sum_{i=1}^{l}2|X_{1i}\|X_{2i}|\right)\\
&= \frac{k_1s_{21}+k_2s_{22}}{k}+\frac{1}{k}\left(\frac{n^2}{k_1}+\frac{m^2}{k_2}-\frac{(n+m)^2}{k}+2\sum_{i=1}^{l}|X_{1i}\|X_{2i}|\right)\\
&= \frac{k_1s_{21}+k_2s_{22}}{k_1+k_2-l}+\frac{1}{k_1+k_2-l}\left(\frac{n^2}{k_1}+\frac{m^2}{k_2}-\frac{(n+m)^2}{k_1+k_2-l}+2\sum_{i=1}^{l}|X_{1i}\|X_{2i}|\right).\square
\end{aligned}
$$

On the basis of Theorems 1 and 2, an algorithm for determining variance s_h of a large-scale hybrid data table is introduced as follows.

Algorithm 1. Determining variance s_h of a large-scale hybrid data table.

Input: A large-scale hybrid data table $S=(U,A)$, $A=A_1\cup A_2$;
Output: Variance s_h of S.
Step 1: Divide S into several sub-tables, denoted by U_1, $U_2,\ldots,U_t$, where $\sum_{i=1}^{t}|U_i|=|U|$ (the sizes of sub-tables can range from several hundred to several thousand).
Step 2: $s_1\leftarrow 0$, $s_2\leftarrow 0$. Compute the variance s_{11} of numeric data and s_{21} of nominal data on U_1, then $s_1\leftarrow s_{11}$, $s_2\leftarrow s_{21}$.
Step 3: $U'\leftarrow U_1$.
for($i=2; i\le t; i++$)
{ Compute the variance s_{1i} of numeric data and s_{2i} of nominal data on U_i;
According to Theorem 1, compute s_1:

$$s_1\leftarrow\frac{(|U'|-1)s_1+(|U_i|-1)s_{1i}}{|U'|+|U_i|-1}+\Delta_1;$$

According to Theorem 2, compute s_2:

$$s_2\leftarrow\frac{k_1s_2+k_{2i}s_{2i}}{k_1+k_{2i}-l}+\Delta_2;$$

$U'\leftarrow U'\cup U_i$.
}
Step 4: $s_h\leftarrow\frac{|A_1|s_1+|A_2|s_2}{|A|}$, return s_h and end.

For the data sets which are too large in scale to be handled, this algorithm is helpful for solving the variance s_h efficiently. Then, an algorithm for determining sample size is introduced as follows.

Algorithm 2. Determining sample size of a large-scale hybrid data table.

Input: A large-scale hybrid data table $S=(U,A)$, $A=A_1\cup A_2$;
Output: Sample size M_1.
Step 1: Compute s_h on U by using Algorithm 1.
Step 2: Determine the margin error E of S according to experience.
Step 3: According to Definition 5, compute $M'=\frac{Z^2\times s_h}{E^2}$.
Step 4: If $M'>0.05|U|$, then compute adjusted sample size $M_1=\frac{M'\times|U|}{M'+|U|+1}$;
else $M_1\leftarrow M'$.
Step 5: Return M_1 and end.

For supervised feature selection, classes information and selected features are closely related. Thus, to better estimate on sub-tables the feature subset of the original table, classes information contained in sub-tables should be closed to that contained in the original table as far as possible. In the process of collecting sub-tables, we set the numbers of classes in a sub-table equal to that of classes in the original table, and the ratio of sample number of each class of a sub-table equal to that of the original table. Besides, there should be some similarities among sub-tables, which make results on sub-tables close to each other relatively and are more convenient for the fusion of feature subset. Hence, in the selection process of sub-tables, we make each sub-table contains some objects that are identical to those in another one. For convenience, above discussion can be summarized as following three strategies:

(1) *Consistency*: the number of classes in sub-tables are equal to that of classes in the original table; and the ratio of sample number of each class of a sub-table is equal to that of the original table (see Fig. 2).
(2) *Transitivity*: each sub-table contains some objects that are identical to those in another one (see Fig. 3).
(3) *Ergodicity*: all objects in the original large-scale table should be selected into sub-tables as far as possible. In other words, the process of selecting sub-tables is stopped until the number of remaining objects is smaller than the size of sub-table.

On the basis of above three strategies, the specific algorithm for selecting sub-tables is introduced in Algorithm 3.

Algorithm 3. An algorithm for selecting sub-tables from a large-scale hybrid data table.

Input: A hybrid data table $S=(U,A)$, and the set of classes (or labels) $\mathcal{C}=\{c_1,c_2,\ldots,c_r\}$;
Output: n sub-tables $S_j=(U_j,A)(j=1,2,\ldots,n)$.
Step 1: Compute the sample size M_1 on U (according to Algorithm 2).
Step 2: Compute the sets of instances included in each class c_i on U, denoted by C_i, and the class proportions $p_i=\frac{|C_i|}{|U|}(i=1,2,\ldots,r)$.
Step 3: Compute the numbers of instances included in each class c_i in sub-tables $m_i=[M_1\times p_i](i=1,2,\ldots,r)$ (function $[\cdot]$ is the rounding function).
Step 4: Select first sub-table S_1 on U, $U_1\leftarrow\varnothing$: for ($i=1$; $i\le r$; $i++$)
{ Select m_i objects from C_i randomly, which is denoted by X;
$U_1\leftarrow U_1\cup X$;
}
Step 5: Select sub-table S_j repeatedly, $j\leftarrow 2$:
Given a threshold α $(0<\alpha<1)$;
while($|U-\bigcup_{k=1}^{j-1}U_k|>M_1$)
{
Step 5.1: Select αM_1 objects from table S_{j-1}:
{
Compute the sets of instances included in each class c_i on U_{j-1}, denoted by C_i',
Select αm_i objects from C_i' $(i=1,2,\ldots,r)$ randomly, which is denoted by X';
$U_j\leftarrow U_j\cup X'$;
}
Step 5.2: $U''=U-\bigcup_{k=1}^{j-1}U_k$, and select $(1-\alpha)M_1$ objects from U'':
{
Compute the sets of instances included in each class c_i on U'', denoted by C_i'',
Select $(1-\alpha)m_i$ objects from C_i'' $(i=1,2,\ldots,r)$ randomly, which is denoted by X'';
$U_j\leftarrow U_j\cup X''$;
}
$j\leftarrow j+1$;
}
Step 6: $n\leftarrow j-1$ and end.

Here are some explanations about Algorithm 3. In Steps 2–3, the algorithm aims to ensure classes information on sub-tables is close to the large-scale one. Besides, because of that m_i are integers, one can get that $\sum_{i=1}^{r}m_i\approx M$. In the process of selecting sub-tables in Step 5, some objects are selected from the existing sub-tables, which ensure there are certain similarities among selected sub-tables. In addition, threshold α should not be too small to weaken the similarity, here we propose an empirical value of $\alpha=0.3$–0.5.

3.2. Feature selection to sub-tables

On the basis of fuzzy rough set model, Hu et al. [10] defined the information entropy based on fuzzy equivalence relation, redefined the feature significance, and proposed a forward search feature selection algorithm for hybrid data sets. To select features from the collected samples, this algorithm is employed to find

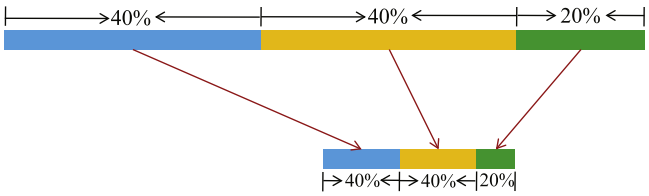

Fig. 2. Consistency.

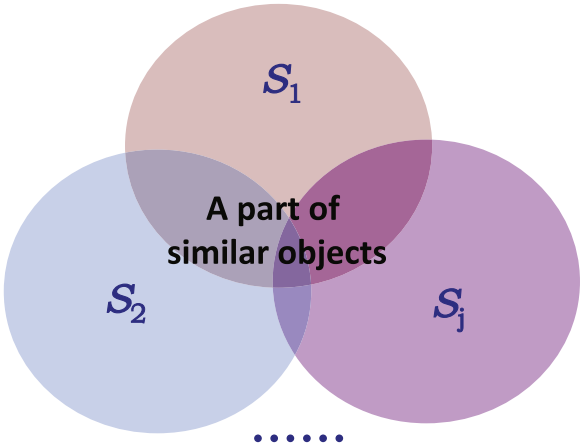

Fig. 3. Transitivity.

feature subset in this paper. The specific algorithm steps are formulated as follows.

Algorithm 4. An algorithm for calculating feature subset.

Input: A hybrid decision table $S=(U,C\cup D)$.
Output: A feature subset *Red* of S.
Step 1: $Red\leftarrow\varnothing$.
Step 2: Compute and select sequentially
$SIG(a_0,Red,D)=\max\{SIG(a_i,Red,D)\},a_i\in C-Red$.
Step 3: If $SIG(a_0,Red,D)>0$, then $Red\leftarrow Red\cup a_0$, goto step 2; else goto step 4;
Step 4: Return *Red* and end.

According to the definition of fuzzy equivalence matrix, its complexity is $O(|C||U|^2)$. Hence, the complexity of Algorithm 4 is $O(|C|^2|U|^2)$.

3.3. Fusion of feature subsets

According to Algorithms 3 and 4, one can obtain a group of estimates of feature selection to a given large-scale hybrid decision table. This section introduces an approach for fusing together all estimates and generating a valid feature subset. In Algorithm 4, one feature with the highest significance is added to a pool at each iteration, thus forming an ordered feature subset. According to the sort, we assign a weight to each feature. Thus,

$$w_i=\frac{2(n-i+1)}{n(n+1)}$$

denotes the weight of ith feature in an ordered feature subset with n features, where

$$w_i=\frac{2(n-i+1)}{n(n+1)}=\frac{n-i+1}{\frac{n(n+1)}{2}}.$$

Obviously, the more significant feature is assigned a bigger weight. Then, the weighted frequency of each feature in estimates is counted, and more frequent feature is considered as more significant. We rank all features in estimates according to their weighted frequencies and generate a final ordered feature subset. In addition, in the

experiments, by using existing well-known classifiers or learning algorithms, we can remove some not very important features from the end of the ordered feature subset and retain the more significant ones. The specific algorithm is formulated as follows.

Algorithm 5. An efficient feature selection algorithm for large-scale hybrid decision tables (EFSH).

Input: A hybrid large-scale decision table $S=(U, C\cup D)$;
Output: A feature subset R.
Step 1: Select n sub-tables by using Algorithm 3 from S:
$S_1=(U_1, C\cup D)$, $S_2=(U_2, C\cup D)$, …, $S_n=(U_n, C\cup D)$.
Step 2: $R\leftarrow\varnothing$, $cou_w(a_i)=0$ $(i=1,2,\ldots,|C|)$.
Step 3: *For* $(j=1; j\le n; j++)$
{
Find the feature subset $red_j=\{a'_1, a'_2, \ldots, a'_{k'}\}$ of S_j by using Algorithm 4;
$\forall a'_i\in red_j$, $cou_w(a'_i)\leftarrow cou_w(a'_i)+\frac{2(k'-i+1)}{k'(k'+1)}$.
}
Step 4: Sort $cou_w(a_i)$ from large to small, denoted by $\{cou_w(a''_1), cou_w(a''_2), \ldots, cou_w(a''_k)\}$ $(k\le|C|)$.
Step 5: return $R=\{a''_1, a''_2, \ldots, a''_k\}$ and end.

Table 1
Description of data sets.

	Data sets	Samples	Features			Classes
			Total	Numeric	Nominal	
1	credit	690	15	6	9	2
2	anneal	898	38	6	32	6
3	vowel	990	13	10	3	11
4	german	1000	20	7	13	2
5	sick	3772	29	7	22	2
6	hypothyroid	3772	29	6	23	4
7	waveform	5000	40	40	0	3
8	Ticdata	5822	85	0	85	2
9	thyroid	9172	29	7	22	2

Table 2
Comparison of computational time.

Data sets	Computational time/s		
	A4	EFSH	PIT (%)
credit	24.03	10.26	57.30
anneal	162.06	59.39	63.35
vowel	41.82	4.66	88.86
german	19.63	8.45	56.95
sick	1516.71	310.06	79.56
hypothyroid	1446.28	148.09	89.76
waveform	4231.96	411.24	90.28
Ticdata	296.38	140.41	52.63
thyroid	15,279.83	1534.34	89.96

Because the time complexity of Algorithm 4 is $O(|C|^2|U|^2)$, the time complexity of Algorithm 5 is $O(|C|^2(|U_1|^2+|U_2|^2+\cdots+|U_n|^2))$. Usually, $|U|^2$ is much larger than $|U_1|^2+|U_2|^2+\cdots+|U_n|^2$. Therefore, the computational time of algorithm EFSH is much smaller than that of Algorithm 4.

It should be noted that this algorithm introduces a framework that is dividing and fusing on a large-scale hybrid data set. Based on this framework, by employing other feature selection algorithms to select features on a sub-table, one can also construct appropriate efficient algorithms.

4. Experimental analysis

The objective of following experiments is to show effectiveness of algorithm EFSH. All the experiments were carried out on a personal computer with Windows 7, Inter(R) Core (TM) i7-2600 CPU (2.66 GHz) and 4.00 GB memory. The software being used is Microsoft Visual Studio 2010 and programming language is C#. Note that, for selected sub-tables, most of the existing algorithms for hybrid data can be employed to select features. We mainly focus on in this paper how to select sub-tables and fuse the estimates. Because in this paper (Section 4.1), Hu's algorithm (Algorithm 4) is employed to do feature selection on sub-tables, here we only compare the performance of the proposed algorithm EFSH and Hu's algorithm.

Twelve UCI data sets are employed to illustrate feasibility and efficiency of algorithm EFSH in this section. In Section 4.1, efficiency of EFSH is illustrated mainly through comparing computational time of algorithm EFSH and Algorithm 4. In Section 4.2, 10-fold cross validation and four classical classification algorithms are employed to evaluate algorithm EFSH and Algorithm 4. In Section 4.3, to further illustrate efficiency, three very large-scale data sets are employed to conduct the experiment. The specific experiments of each part is introduced as follows.

4.1. Efficiency analysis

In this subsection, nine UCI data sets shown in Table 1 are employed to test algorithm EFSH and Algorithm 4. The feature selection results of the two algorithms are showed Tables 2 and 3. Table 2 shows the comparison of computational time. Table 3 shows the selected features. In Table 2, " Algorithm 4" is simplified as A4 and "*percentage improvement of computational time*" is simplified as PIT.

Experimental results in Table 2 show that computational time of EFSH is much shorter than that of Algorithm 4. Because of that the main idea of EFSH is decomposing a large-scale data set into a family of small ones and doing feature selection on these small data sets. Obviously, the total time spent on selecting features from small data sets is much less than that for the original large-scale one, EFSH yields in a much less amount of time a feature subset. Results in Table 2 well validate this conclusion. In Table 2,

Table 3
Comparison of selected features.

Data sets	A4	EFSH
credit	1,2,3,4,6,7,8,9,10,11,12,13,14,15	1,2,3,4,6,7,8,9,10,11,12,13,14
anneal	1,3,4,5,7,8,9,13,31,33,34,35	1,3,4,5,7,8,9,31,33,34,35,36
vowel	1,3,6,7,9,11,12,13	4,5,6,7,8,9,11,12,13
german	3,4,5,6,7,9,11,12,16	3,4,5,6,7,9,11,12
sick	1,3,4,6,8,24,29	1,2,3,4,25,29
hypothyroid	1,15,17,18,20,22	1,16,18,19,20,22,26
waveform	2,7,11,12,13,15,16	1,2,3,7,10,11,12,15
Ticdata	2,5,7,15,17,31,38,43,44,45,47,48,49,54,55,57,58,59,61,63,64,68,80,83	2,3,5,7,8,15,17,18,19,30,31,39,43,44,45,47,48,49,52,54,55,57,59,61,62,64,68,80,83
thyroid	1,2,3,4,6,7,23,24,25,26,28	1,2,3,4,6,8,20,23,24,25,26

Table 4
Classification accuracies of A4.

Data sets	*N*	NBC	C4.5	JRip	RF
credit	14	0.7724 ± 0.2256	**0.8608 ± 0.1923**	0.8463 ± 0.2278	0.8333 ± 0.2361
anneal	11	**0.7973 ± 0.0782**	0.9354 ± 0.0274	0.9376 ± 0.0204	0.9387 ± 0.0250
vowel	9	0.3292 ± 0.1412	0.6323 ± 0.0702	0.4949 ± 0.1040	0.8090 ± 0.0710
german	9	0.7010 ± 0.3881	**0.6920 ± 0.4109**	0.689 ± 0.4109	0.6480 ± 0.3848
sick	7	**0.9528 ± 0.0711**	0.9835 ± 0.0231	**0.9379 ± 0.1047**	**0.9273 ± 0.0899**
hypothyroid	7	0.9456 ± 0.0339	0.9734 ± 0.0184	0.9729 ± 0.0198	0.9689 ± 0.0187
waveform	15	**0.7846 ± 0.1703**	0.7550 ± 0.1974	**0.7690 ± 0.2190**	**0.7774 ± 0.1830**
Ticdata	28	**0.8885 ± 0.1359**	0.9402 ± 0.1124	0.9397 ± 0.1121	**0.9266 ± 0.1051**
thyroid	11	0.9323 ± 0.0212	0.9371 ± 0.0560	0.9326 ± 0.0581	0.9297 ± 0.0528

Table 5
Classification accuracies of EFSH.

Data sets	*N*	NBC	C4.5	JRip	RF
credit	13	**0.8130 ± 0.1954**	0.8522 ± 0.2027	**0.8579 ± 0.2261**	**0.8492 ± 0.2231**
anneal	11	0.7109 ± 0.0964	**0.9555 ± 0.0171**	**0.9577 ± 0.0229**	**0.9621 ± 0.0188**
vowel	9	**0.6456 ± 0.0802**	**0.7900 ± 0.0367**	**0.6789 ± 0.0628**	**0.9211 ± 0.0446**
german	8	**0.7080 ± 0.3691**	0.6870 ± 0.4170	**0.6870 ± 0.4082**	**0.6710 ± 0.3843**
sick	6	0.9483 ± 0.0801	**0.9857 ± 0.0273**	0.9377 ± 0.1153	0.9247 ± 0.1042
hypothyroid	7	**0.9477 ± 0.0386**	**0.9738 ± 0.0207**	**0.9753 ± 0.0174**	0.9592 ± 0.0233
waveform	15	0.7828 ± 0.1758	**0.7674 ± 0.1974**	0.7580 ± 0.2286	0.7624 ± 0.1741
Ticdata	29	0.8829 ± 0.1424	0.9402 ± 0.1124	0.9397 ± 0.1121	0.9241 ± 0.1079
thyroid	11	**0.9353 ± 0.0612**	**0.9379 ± 0.0559**	**0.9361 ± 0.0565**	**0.9305 ± 0.0523**

Table 6
Description of data sets for high-efficiency.

	Data sets	Samples	Features			Classes
			Total	Numeric	Nominal	
1	census-income	299,285	40	7	33	2
2	kddcup	4,898,431	41	34	7	23

compared with computational time of A4, EFSH saves more than 50% computational time of all of employed data sets. Particularly for data sets *hypothyroid*, *waveform* and *thyroid*, EFSH saves nearly 90% computational time. Hence, algorithm EFSH can get a feature subset in very short time when dealing with the relatively large-scale data sets. Table 3 shows that most of the best features selected by using EFSH and A4 are the same. In view of the fusing mechanism used in EFSH, there are some difference in the feature subsets of EFSH and A4 is unavoidable. This conclusion is well validated in the experiments shown in Table 3. In short, results in this subsection show that EFSH can find a very similar feature subset with A4 in a much shorter time. Specifically, the accuracies of feature selection results are compared in next subsection.

4.2. Effectiveness analysis

In this subsection, four classical classifiers are employed to evaluate the feature selection results shown in Table 3. The four classifiers are Naive Bayes classifier, C4.5 classifier, JRip classifier and RandomForest classifier. Based on the classifiers, Table 4 shows classification accuracies of feature subsets selected by using A4, and Table 5 shows classification accuracies of feature subsets of EFSH. In Tables 4 and 5, "*NaiveBayes*" is simplified as NBC and "*RandomForest*" is simplified as RF, "*N*" is the number of selected features.

From the results in Tables 4 and 5, one can easily see that, for each classifier and each data set, classification accuracies based on the two algorithms are very close to each other. Specifically, for Naive Bayes classifier (NBC), there are five data sets whose classification accuracies based on EFSH are higher than that based on A4. For C4.5 classifier, there are six data sets based on EFSH get higher accuracies. For JRip classifier and RandomForest classifier, there are respectively six data sets and five data sets based on EFSH get higher accuracies. For convenience, the values of higher accuracies are bold in Tables 4 and 5. In summary, the performance of the feature subsets found by the two algorithms are very close to each other without obvious superiority and inferiority. In other words, compared with A4, EFSH can find an effective feature subset, whereas EFSH saves much more computational time. From a combination of results in Sections 4.1 and 4.2, one can get that, using the idea of decomposition and fusing to deal with large-scale data sets is effective and feasible. By using the idea, algorithm EFSH finds an effective feature subset on the condition of saving a lot of computational time.

4.3. Efficiency analysis for very large-scale data sets

Experimental results in Sections 4.1 and 4.2 show that *EFSH* can find an effective feature subset in a much shorter time. To further demonstrate its efficiency, two UCI very larger-scale data sets outlined in Table 6 are employed to conduct the experiments in this subsection. By using some representative algorithms for hybrid data and Algorithm 4, these two data sets are too large in scale to get a feature subset within 150 h on a PC. In this section, we try to carry out the proposed algorithm EFSH on them and find an effective feature selection result. The experimental results are shown in Table 7. In addition, because of that we did not get the results by using existing feature selection algorithms for hybrid data, Table 7 just shows the results of EFSH. Two classic classifiers Naive Bayes classifier (NBC) and C4.5 are introduced to evaluate the selected features in this subsection.

The experimental results shown in Table 7 indicate that, for the two data sets, the new proposed algorithm EFSH can find their feature subsets within just 4.38 h and 33.88 h on a PC, respectively. Moreover, the accuracies of EFSH are same to, or even higher than that of the raw data. As mentioned above,

Table 7
Classification accuracies of selected features.

Data sets	Raw data		EFSH			
	NBC	C4.5	N	NBC	C4.5	Time/h
census-income	0.8583 ± 0.1442	0.9507 ± 0.0798	23	0.9083 ± 0.0997	0.9506 ± 0.0800	4.38
kddcup	0.9948 ± 0.0005	0.9996 ± 0.0001	21	0.9950 ± 0.0005	0.9997 ± 0.0000	33.88

experiments in this subsection is just try to deal with the very large-scale data sets which cannot be handled in an efficient way by using other feature selection algorithms. Hence, this subsection does not give more comparison with other algorithms. In summary, experiments in Table 7 well validate the efficiency of EFSH, especially for large-scale data sets.

5. Conclusions

Feature selection is a significant dimensional reduction technique in machine learning. For large-scale hybrid data sets, by decomposing large data sets and fusing results of sub-tables, an efficient algorithm for selecting informative features has been proposed in this paper. Experiments show that the algorithm is effective and efficient, especially for large-scale data sets. Note that the new algorithm not only saves much more computational time, but also can deal with the large-scale data sets which are very difficult to handle because of the high computational time. It is our wish that this study provides new views and thoughts on exploring efficient machine learning approaches for big data.

Acknowledgment

This work was supported by National Natural Science Fund of China (Nos. 61402272, 71301090, 61303008), National Key Basic Research and Development Program of China (973) (No. 2013CB329404).

References

[1] H. Almuallim, T.G. Dietterich, Learning boolean concepts in the presence of many irrelevant features, Artif. Intell. 69 (1–2) (1994) 279–305.
[2] K. Benabdeslem, M. Hindawi, Efficient semi-supervised feature selection: constraint, relevance and redundancy, IEEE Trans. Knowl. Data Eng. 26 (2004) 1131–1143.
[3] M. Dash, H. Liu, Feature selection for classification, Intell. Data Anal. 1 (1997) 131–156.
[4] M. Dash, H. Liu, Consistency-based search in feature selection, Artif. Intell. 151 (2003) 155–176.
[5] M. Dash, H. Liu, J. Yao, Dimensionality reduction of unsupervised data, in: Proceedings of the Ninth IEEE International Conference on Tools with AI (ICTAI 97), 1997, pp. 532–539.
[6] I. Guyon, A. Elisseeff, An introduction to variable feature selection, Mach. Learn. Res. 3 (2003) 1157–1182.
[7] X.H. Hu, N. Cercone, Learning in relational databases: a rough set approach, Int. J. Comput. Intell. 11 (2) (1995) 323–338.
[8] M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in: Proceedings of the 17th International Conference on Machine Learning, 2000, pp. 359–366.
[9] Q.H. Hu, Z.X. Xie, D.R. Yu, Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation, Pattern Recognit. 40 (2007) 3509–3521.
[10] Q.H. Hu, D.R. Yu, Z.X. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognit. Lett. 27 (5) (2006) 414–423.
[11] Q.H. Hu, D.R. Yu, J.F. Liu, C.X. Wu, Neighborhood rough set based heterogeneous feature subset selection, Inf. Sci. 178 (2008) 3577–3594.
[12] V.N. Huynh, Y. Nakamori, A roughness measure for fuzzy sets, Inf. Sci. 173 (2005) 255–275.
[13] R. Jensen, Q. Shen, Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches, IEEE Trans. Knowl. Data Eng. 16 (12) (2004) 1457–1471.
[14] R. Jensen, Q. Shen, Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches, IEEE Press/Wiley & Sons, Canada, 2008.
[15] J.P. Jia, Principles of Statistics, fourth ed., Renmin University Publishing, Beijing, China, 2009.
[16] F. Jiang, Y.F. Sui, C.G. Cao, Some issues about outlier detection in rough set theory, Expert Syst. Appl. 36 (3) (2009) 4680–4687.
[17] Y.J. Jin, Z.F. Du, Y. Jiang, Sampling Technique, China Renmin University Publishing, Beijing, 2008.
[18] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1–2) (1997) 273–324.
[19] K. Kira, L.A. Rendell, The feature selection problem: traditional methods and a new algorithm, Proc. AAAI 92 (1992) 129–134.
[20] M. Kryszkiewicz, P. Lasek, FUN: fast discovery of minimal sets of attributes functionally determining a decision attribute, Trans. Rough Sets 9 (2008) 76–95.
[21] N. Kwak, C.H. Choi, Input feature selection by mutual information based on Parzen window, IEEE Trans. Pattern Anal. Mach. Intell. 24 (12) (2002) 1667–1671.
[22] C.K. Lee, G.G. Lee, Information gain and divergence-based feature selection for machine learning-based text categorization, Inf. Process. Manag. 42 (2006) 155–165.
[23] H. Liu, H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Kluwer Academic, Boston, 1998.
[24] H. Liu, F. Hussain, M. Dash, Discretization: an enabling technique, Data Min. Knowl. Discov. 6 (4) (2002) 393–423.
[25] H. Liu, L. Yu, Toward integrating feature selection algorithms for classification and clustering, IEEE Trans. Knowl. Data Eng. 17 (4) (2005) 491–502.
[26] T.R. Li, D. Ruan, W. Geert, J. Song, Y. Xu, A rough sets based characteristic relation approach for dynamic attribute generalization in data mining, Knowl.-Based Syst. 20 (5) (2007) 485–494.
[27] D. Liu, T.R. Li, D. Ruan, W.L. Zou, An incremental approach for inducing knowledge from dynamic information systems, Fundam. Inform. 94 (2009) 245–260.
[28] J.Y. Liang, F. Wang, C.Y. Dang, Y.H. Qian, A group incremental approach to feature selection applying rough set technique, IEEE Trans. Knowl. Data Eng. 26 (2) (2014) 294–308.
[29] J.Y. Liang, F. Wang, C.Y. Dang, Y.H. Qian, An efficient rough feature selection algorithm with a multi-granulation view, Int. J. Approx. Reason. 53 (2012) 912–926.
[30] J.Y. Liang, X.W. Zhao, D.Y. Li, F.Y. Cao, C.Y. Dang, Determining the number of clusters using information entropy for mixed data, Pattern Recognit. 45 (6) (2012) 2251–2265.
[31] Z.Q. Meng, Z.Z. Shi, A fast approach to attribute reduction in incomplete decision systems with tolerance relation based rough sets, Inf. Sci. 179 (2009) 2774–2793.
[32] W. Pedrycz, G. Vukovich, Feature analysis through information granulation and fuzzy sets, Pattern Recognit. 35 (2002) 825–834.
[33] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Boston, 1991.
[34] Z. Pawlak, Rough set theory and its applications in data analysis, Cybern. Syst. 29 (1998) 661–688.
[35] Z. Pawlak, A. Skowron, Rough sets and boolean reasoning, Inf. Sci. 177 (1) (2007) 41–73.
[36] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[37] Y.H. Qian, J.Y. Liang, W. Pedrycz, C.Y. Dang, Positive approximation: an accelerator for attribute reduction in rough set theory, Artif. Intell. 174 (2010) 597–618.
[38] R.W. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognit. Lett. 24 (2003) 833–849.
[39] W.H. Shu, H. Shen, Incremental feature selection based on rough set in dynamic incomplete data, Pattern Recognit. 47 (2014) 3890–3906.
[40] M.W. Shao, W.X. Zhang, Dominance relation and rules in an incomplete ordered information system, Int. J. Intell. Syst. 20 (2005) 13–27.
[41] W.Y. Tang, K.Z. Mao, Feature selection algorithm for mixed data with both nominal and continuous features, Pattern Recognit. Lett. 28 (5) (2007) 563–571.
[42] R. Thawonmas, S. Abe, A novel approach to feature selection based on analysis of class regions, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 27 (2) (1997) 196–207.

[43] N.X. Vinh, J. Bailey, Comments on supervised feature selection by clustering using conditional mutual information-based distances, Pattern Recognit. 46 (4) (2013) 1220–1225.
[44] H. Wang, D. Bell, F. Murtagh, Axiomatic approach to feature subset selection based on relevance, IEEE Trans. Pattern Anal. Mach. Intell. 21 (3) (1999) 271–277.
[45] F. Wang, J.Y. Liang, Y.H. Qian, Attribute reduction: a dimension incremental strategy, Knowl.-Based Syst. 39 (2013) 95–108.
[46] F. Wang, J.Y. Liang, C.Y. Dang, Attribute reduction for dynamic data sets, Appl. Soft Comput. 13 (2013) 676–689.
[47] J.Z. Wang, L.S. Wu, J. Kong, Y.X. Li, B.X. Zhang, Maximum weight and minimum redundancy: a novel framework for feature subset selection, Pattern Recognit. 46 (6) (2014) 1616–1627.
[48] H. Wang, Nearest neighbors by neighborhood counting, IEEE Trans. Pattern Anal. Mach. Intell. 28 (6) (2006) 942–953.
[49] W. Wei, J.Y. Liang, Y.H. Qian, F. Wang, An attribute reduction approach and its accelerated version for hybrid data, in: The 8th IEEE International Conference on Cognitive Informatics, 2009, pp. 167–173.
[50] W.Z. Wu, J.S. Mi, W.X. Zhang, Generalized fuzzy rough sets, Inf. Sci. 151 (2003) 263–282.
[51] Z.Y. Xu, Z.P. Liu, B.R. Yang, W. Song, A quick attribute reduction algorithm with complexity of max $(O(|C||U|), O(|C|^2|U/C|))$, Chin. J. Comput. 29 (3) (2006) 391–398.
[52] W.H. Xu, X.Y. Zhang, W.X. Zhang, Knowledge granulation, knowledge entropy and knowledge uncertainty measure in ordered information systems, Appl. Soft Comput. 9 (4) (2009) 1244–1251.
[53] Y.Y. Yao, Neighborhood systems and approximate retrieval, Inf. Sci. 176 (23) (2006) 3431–3452.
[54] Y.Y. Yao, Y. Zhao, Attribute reduction in decision-theoretic rough set models, Inf. Sci. 178 (17) (2008) 3356–3373.
[55] L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learn. Res. 5 (2004) 1205–1224.

Feng Wang works in the School of Computer and Information Technology, Shanxi University. She received the PhD degree in Computers with applications at Shanxi University (2013). Her research interests are in the areas of feature selection, rough set theory, artificial intelligence.

Jiye Liang is a professor of School of Computer and Information Technology and Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education at Shanxi University. He received the PhD degree in Applied Mathematics from Xi'an Jiaotong University. His research interests include feature selection, artificial intelligence, granular computing, data mining and knowledge discovery. He has published more than 80 articles in international journals.