Neurocomputing 168 (2015) 210–220

Contents lists available at ScienceDirect

Neurocomputing

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

Mutual information criterion for feature selection from incomplete data

Wenbin Qian [a,b,*], Wenhao Shu [c]

[a] School of Software, Jiangxi Agriculture University, Nanchang 330045, China
[b] Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China
[c] School of Information Engineering, East China Jiaotong University, Nanchang 330013, China

ARTICLE INFO

Article history:
Received 23 July 2014
Received in revised form
2 April 2015
Accepted 28 May 2015
Communicated by Feiping Nie
Available online 17 June 2015

Keywords:
Feature selection
Uncertainty measure
Mutual information
Incomplete data
Rough sets

ABSTRACT

Feature selection is an important preprocessing step in machine learning and data mining, and feature criterion arises a key issue in the construction of feature selection algorithms. Mutual information is one of the widely used criteria in feature selection, which determines the relevance between features and target classes. Some mutual information-based feature selection algorithms have been extensively studied, but less effort has been made to investigate the feature selection issue in incomplete data. In this paper, combined with the tolerance information granules in rough sets, the mutual information criterion is provided for evaluating candidate features in incomplete data, which not only utilizes the largest mutual information with the target class but also takes into consideration the redundancy between selected features. We first validate the feasibility of the mutual information. Then an effective mutual information-based feature selection algorithm with forward greedy strategy is developed in incomplete data. To further accelerate the feature selection process, the selection of candidate features is implemented in a dwindling object set. Compared with existing feature selection algorithms, the experimental results on different real data sets show that the proposed algorithm is more effective for feature selection in incomplete data at most cases.

1. Introduction

In machine learning and data mining, one is often confronted with the data sets involving huge number of features or even overwhelming feature set. Many redundant or irrelevant features may lead to poor performance of the learning algorithms, and increase in the size of the search space when using data mining tools for knowledge discovery. To mitigate this problem, one effective way to reduce the dimensionality of feature space is the feature selection technique. The main goal of feature selection is to find a feature subset that has the most discriminative information from the original feature set. In many real-world applications, feature selection has shown to be very effective in reducing dimensionality, removing irrelevant and redundant features, improving the learning accuracy, efficiency and scalability of a classification task, and enhancing learning comprehensibility [2,5,17,19,20,29,44].

According to whether the evaluation measure is available, feature selection methods can broadly fall into the filter and the wrapper methods [22,41]. Filter methods evaluate the goodness of a feature or set of features by using only intrinsic characteristics of the data, which is independent of any learning algorithm, while wrapper methods rely on the performance of a predefined learning algorithm to evaluate the goodness of a subset of features [22]. For high-dimensional data, filter methods are often preferred due to their computational efficiency. To data, many efficient filter feature selection algorithms have been proposed in the literature. Typically, there are two issues in constructing the filter feature selection algorithms: feature evaluation and search strategy. Feature evaluation is used to measure the quality of candidate features. So far, many evaluation criteria are designed, such as dependency [26,42], mutual information [8,24], sample margin [27] in statistical learning theory, and so on. It is noteworthy that among various evaluation criteria, mutual information, as an effective metric to scale the relevance between features, achieves excellent performance and has drawn more and more attention. As to the search strategy, it can be roughly divided into two main categories. One is to find the optimal feature subset in terms of the evaluation criteria, such as the exhaustive search [29] and the branch-and-bound search [30]. The other is to find a suboptimal solution for efficiency, including sequential forward selection [31], sequential backward elimination [33], floating search [32], and so

* Corresponding author.
E-mail addresses: qianwenbin1027@126.com (W. Qian), 11112084@bjtu.edu.cn (W. Shu).

http://dx.doi.org/10.1016/j.neucom.2015.05.105
0925-2312/

on. Feature selection aims at selecting the optimal feature subset based on certain search strategy. However, it is essentially a combinatorial optimization problem, which is computationally expensive [26]. Given a feature set size n, the feature selection is the task of searching for an optimal feature subset through competing 2^n candidate subsets. Although an exhaustive method may be used for this purpose, this is quite impractical for most large-scale data sets. Consequently, to avoid this prohibitive complexity, we employ the sequential forward search strategy to perform the feature selection in this paper.

Granular computing (GrC) has attracted much attention in recent years. The basic idea in GrC is to employ granules and relationships between granules to find solutions [36,46,48]. Three examples of granular computing theories are rough set theory [11], fuzzy logic theory [28], and quotient space theory of problem solving [21]. It is worth mentioning that rough set theory, proposed by Pawlak, is a powerful mathematical tool for data mining, feature selection, intelligent data analysis, decision making and so on [4,9,16,18,26,45,49]. The main advantage of rough set theory is that it can be used as a tool to reduce the number of features contained in data sets using the supplied data alone, requiring no other information, while most other theories require supplementary knowledge, such as Dempster–Shafer theory that requires probability assignments [43] and fuzzy set theory that requires membership values [25]. Information granulation is an important concept in the rough set theory. A granule is a clump of objects drawn together by indistinguishability, similarity, and proximity of functionality [36]. Granulation of an object leads to a collection of granules. In classical rough sets, the granulation of objects induced by an equivalent relation is a set of equivalence classes, in which each equivalence class can be regarded as an (Pawlak) information granule [11,38]. However, the classical rough sets, based on equivalence relation, can only deal with complete data sets, whereas incomplete data sets with missing feature values, more common in real world, are beyond its scope. Thus the Pawlak information granule is not suitable for incomplete data. To be applicable for incomplete data, two main kinds of information granules are generated as follows [12,40]: tolerance information granules and similarity information granules. Induced by a tolerance relation, the granulation of objects generates a set of tolerance classes, in which each tolerance class can be seen as a tolerance information granule. Induced by a similarity relation, the granulation of objects generates a set of similarity classes, in which each similarity class can be seen as a similarity information granule [40]. The way in which the granulation of objects generated by a family of the tolerance information granules is representative and extensively studied [3,9,12,16,18,34,35].Thus the proposed work is focused on a family of tolerance information granules for incomplete data.

Rough set-based data analysis starts from a data table, called an information system. The information system contains data about objects of interest characterized by a finite set of features. If condition features and decision features in an information system are distinguished, it is called a decision system. Although many feature selection algorithms have been proposed for complete decision systems, they have their inability to handle incomplete decision systems (some feature values are missing). But in many practical applications, it may happen that some data sets are usually corrupted by noise, and the feature values for some objects are missing [9,15,18,34,35]. For example, data are collected on patients sick with flu and one of the features is age, with specific age values. Some patients may feel that this feature involves with their privacies, and they refuse to answer. The classical rough set theory defined with equivalence relations leads to the limitation in handling data with missing features. However, feature selection is clearly desirable due to the abundance of missing features in many real-world applications. To avoid information loss, we can select a subset of features by rough sets although some features are missing, and preserve the meaning of features contained in the data set. In the feature selection process, mutual information is an effective metric to scale the relevance between features. It has been applied in diverse fields as a very useful mechanism for characterizing information contents [6–8,13]. With the merits of the two methods, some researchers have employed the combination of mutual information and rough sets for feature selection. Xu et al. [14] developed a mutual information-based algorithm for feature selection based on fuzzy rough set theory and information theory in the fuzzy decision table. Zhou et al. [39] proposed a feature selection algorithm based on mutual information and rough sets for microarray data. Unfortunately, few studies have addressed the feature selection issue in incomplete data by combining the mutual information and information granules based on rough sets, thus we focus on this issue in this paper.

The paper is organized as follows. Section 2 briefly reviews some related work on feature selection involved in this paper. Section 3 introduces some basic concepts. Section 4 develops a mutual information-based feature selection algorithm from incomplete data. Sections 5 shows extensive experiments to demonstrate the effectiveness and efficiency of the proposed algorithm, compared with several existing methods on different data sets. Finally, Sections 6 presents the conclusions and future work.

2. Related work

In this section, we briefly summarize some related work on feature selection in incomplete data and review some existing mutual information-based feature selection algorithms.

Based on rough set theory, some methods for feature selection from incomplete decision systems have been proposed. Accordingly, the methods can be divided into three main categories: one based on positive region [9,18,34], the other based on discernibility matrix [12,35,49], another based on entropy [3,10,16]. Each of these methods preserves a particular property of a given incomplete decision system. Applying the method of positive region, Qian et al. [9] provided a technique called positive approximation to perform efficient feature selection from incomplete decision systems. Parthalain et al. [18] used the information contained within the boundary region to guide the feature selection process under the tolerance rough set model. Meng et al. [34] developed a positive region-based algorithm to solve the feature selection problem in incomplete decision systems. By using the positive region-based approaches, the certain knowledge hidden in data can only be extracted. From the viewpoint of discernibility matrix, Kryszkiewicz [12] gave a discernibility matrix-based method for feature selection in incomplete information systems, in which any two objects determine one feature subset that can distinguish them. By the discernibility matrix-based method, Qian et al. [35] constructed two discernibility matrices associated with two approximation feature selections for inconsistent incomplete decision tables. Though the above discernibility matrix methods can find all the subsets of features in incomplete decision systems, they are usually time-consuming and intolerable to process large-scale incomplete data. To reduce the storage space of the existing discernibility matrix-based feature selection methods, Yang et al. [49] developed a feature selection algorithm based on a novel condensing tree structure (*C*-Tree). From the perspective of entropy, Sun et al. [3] proposed a rough entropy-based feature selection algorithm by using the uncertainty measure in incomplete decision systems. Dai et al. [16] introduced a new uncertainty measure to evaluate the uncertainty of an incomplete decision

table, and applied the new conditional entropy to reduce redundant features in incomplete decision systems. By using the entropy-based approaches, the uncertain knowledge hidden in data can only be extracted. The above shows that few studies take into consideration the mutual information with rough sets for feature selection in incomplete data.

Mutual information in Shannon's information theory is considered as a good indicator of relevance between two random variables [21]. Since mutual information is good at quantifying how much information is shared by two random variables, it is often taken as evaluation criterion to measure the relevance between features and the target classes. Recently, efforts have done to adopt the mutual information in the feature selection process for classification tasks [6–8,13,23,24,47]. Kwak and Choi [8] proposed a method of calculating mutual information based on Parzen window, and developed a mutual information-based feature selection algorithm. However, they do not take into consideration how the selected features work together in classification. Jain and Zongker [5] proved that the combinations of individually good features do not necessarily lead to good classification performance. Liu et al. [23] proposed a feature selection algorithm based on dynamic mutual information, which is estimated on the unlabeled objects. However, the proposed feature selection algorithm is applicable to complete data. Peng et al. [6] presented a two-stage feature selection algorithm by combining minimal-redundancy-maximal-relevance (mRMR) and wrappers. Huang and Chow [7] proposed a mutual information-based feature selection scheme by integrating into a supervised data compression algorithm. Estevez et al. [13] introduced the feature criterion called normalized mutual information for selecting a feature subset. The feature selection scheme ignores the interactive effect of the incumbent feature with those features yet to be identified. Yang and Ong [24] proposed a feature selection algorithm using a mutual information-based feature criterion that measures the importance of features in a backward selection framework. However, the joint effect of features on the target class in the classification task has not been considered, since the selected features may have the redundancy. From the above, few studies have investigated the feature selection issue in incomplete data from the perspective of the mutual information combined with the tolerance information granules in rough sets, which motivates our study in this paper. In our proposed work, the mutual information criterion is provided for evaluating the quality of candidate features, which not only utilizes the largest mutual information with the target class but also takes into consideration the redundancy between selected features. Moreover, to accelerate the feature selection process, the selection of candidate features is implemented in a dwindling object set.

3. Basic concepts

In this section, we will introduce some basic concepts involved in this paper, such as the incomplete decision system, tolerance relation and mutual information.

3.1. Incomplete data

Data sets are usually given as the form of tables, we call a data table as an information system, formulated as $IS=\langle U,A,V,f\rangle$, where U is a set of nonempty and finite objects, called the universe; A is the set of features characterizing the objects; V is the union of feature domains, i.e., $V=\cup_{a\in A}V_a$, where V_a is the value set of feature a, called the domain of a; and $f: U\times A\to V$ is an information function, which assigns feature values to objects such that $\forall a\in A$, $x\in U$, and $f(x,a)\in V_a$, where $f(x,a)$ denotes the value of feature a for object x. If the feature set A is divided into condition feature set C and decision feature set D, then the information system is called a decision system. If there exist $x\in U$ and $a\in A$ such that $f(x,a)$ is equal to a missing value (a null or unknown value, denoted as "$*$"), i.e., $*\in V_a$, then the information system is an incomplete information system (IIS). Otherwise, the information system is a complete information system (CIS). If $*\notin V_D$ but $*\in V_C$, then we call the decision system as an *incomplete decision system* (IDS). If $*\notin V_D$ but $*\notin V_C$, then the decision system is a complete decision system (CDS). Rough set-based feature selection in incomplete data starts from an incomplete decision system.

Given a complete information system $CIS=\langle U,A,V,f\rangle$, for $B\subseteq A$, there is a binary relation $IND(B)$ on U, called the equivalence relation generated by B, defined by $IND(B)=\{(x,y)|\ \forall a\in B, f(x,a)=f(y,a)\}$. Obviously, $IND(B)$ is reflexive, symmetric and transitive. The family of all equivalence classes with respect to $IND(B)$ is denoted as $U/IND(B)$, which is the partition induced by B. An equivalence class of $IND(B)$ containing x is denoted by $[x]_B$. From the viewpoint of granular computing, an equivalence class is also called the Pawlak information granule. If there are missing values, the equivalence relation $IND(B)$ is not suitable for incomplete information systems. Thus Kryszkiewicz defined a kind of tolerance relation as follows [12].

Given an incomplete decision system $IDS=\langle U,A=C\cup D,V,f\rangle$, for $B\subseteq C$, let $TR(B)$ denote the *tolerance relation* between objects that are possibly indiscernible in terms of B, defined by $TR(B)=\{(x,y)|\ \forall a\in B, f(x,a)=f(y,a)\vee f(x,a)=*\vee f(y,a)=*\}$. Obviously, $TR(B)$ is reflexive and symmetric but not transitive. It can be easily shown that $TR(B)=\cap_{a\in B}TR(\{a\})$. The *tolerance information granule* of object x with respect to B is denoted as $T_B(x)=\{y|\ (x,y)\in TR(B)\}$. Let $U/TR(B)$ denote the family set $\{T_B(x)|\ x\in U\}$, which is the classification induced by B.

3.2. Mutual information

From the viewpoint of information granulation, we will present the definition of mutual information for evaluating candidate features in incomplete decision systems. With the mutual information, the purpose of feature selection can find a feature subset that jointly has the relevance on the target classes. In the following, we first give two related definitions of information entropy and conditional entropy in incomplete decision systems.

Definition 1. Let $IDS=\langle U,A=C\cup D,V,f\rangle$ be an incomplete decision system, $U/IND(D)=\{D_1,D_2,\ldots,D_m\}$ is a partition induced by the decision feature set D, the information entropy $H(D)$ of D is defined by $H(D)=-\sum_{i=1}^{m}\frac{|D_i|}{|U|}\log\frac{|D_i|}{|U|}$.

The information entropy is a measure of the information content, which is the uncertainty about knowledge D. Based on the tolerance information granules, the conditional entropy in an incomplete decision system is proposed as follows [16].

Definition 2. Let $IDS=\langle U,A=C\cup D,V,f\rangle$ be an incomplete decision system, $U=\{x_1,x_2,\ldots,x_n\}$, for $B\subseteq C$, the classification induced by B is $U/TR(B)=\{T_B(x_1),T_B(x_2),\ldots,T_B(x_n)\}$, and $U/IND(D)=\{D_1,D_2,\ldots,D_m\}$ is a partition on the decision feature set D, the conditional entropy $H(D|B)$ of D given by B is defined by

$$H(D|B)=-\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{|T_B(x_j)\cap D_i|}{|U|}\log\frac{|T_B(x_j)\cap D_i|}{|T_B(x_j)|}$$

The condition entropy measures the additional amount of information provided by D if B is known. After defining the

concepts of information entropy and condition entropy, then the mutual information is defined as follows.

Definition 3. Let $IDS=\langle U, A=C\cup D, V, f\rangle$ be an incomplete decision system, for $B\subseteq C$, the *mutual information* $I(D;B)$ of D about B is defined by $I(D;B)=H(D)-H(D|B)$.

Mutual information measures the decrease of uncertainty about D caused by B. This measure may be helpful to evaluate the relevance between condition features and target classes. By Definition 3, the significance measure of a feature is given as follows.

Definition 4. Let $IDS=\langle U, A=C\cup D, V, f\rangle$ be an incomplete decision system, for $B\subseteq C$, and $\forall b\in B$, the significance of b in B relative to D is defined by $sig(b,B,D)=I(D;B)-I(D;B-\{b\})$.

From Definition 4, $sig(b,B,D)$ reflects the change of the mutual information if feature b is eliminated from B. The higher the change in mutual information, the more significant the feature is. If $sig(b,B,D)=0$, then the feature b is dispensable. Accordingly, $sig(b,B,D)$ is applicable to select the indispensable features from the whole feature set in the feature selection process.

4. Mutual information criterion for feature selection from incomplete data

In this section, the validity of the mutual information in incomplete data is firstly verified. On this basis, a feature evaluation function is given to find candidate features not only by making use of the largest mutual information with the target classes but also taking into account the redundancy between selected features. And then an efficient mutual information-based feature selection algorithm with greedy forward search strategy is developed from incomplete data. In addition, to further improve the feature selection process, the selection of candidate features is implemented on a dwindling object set.

4.1. Validity of mutual information

In the following, the validity of mutual information will be proved by the theorem.

Theorem 1. *(Monotonicity) Let $IDS=\langle U, A=C\cup D, V, f\rangle$ be an incomplete decision system, for $\forall A, B\subseteq C$, if $A\subseteq B$, then $I(D;A)\le I(D;B)$.*

Proof. Suppose the object set in the incomplete decision system be $U=\{x_1, x_2, \ldots, x_n\}$, the classification of U induced by A is $U/TR(A)=\{T_A(x_1), T_A(x_2), \ldots, T_A(x_n)\}$, the classification induced by B is $U/TR(B)=\{T_B(x_1), T_B(x_2), \ldots, T_B(x_n)\}$, and the partition induced by D is $U/IND(D)=\{D_1, D_2, \ldots, D_m\}$. Since $A\subseteq B$, for $\forall x_j (1\le j\le n)$, it holds that $T_B(x_j)\subseteq T_A(x_j)$, i.e., $|T_B(x_j)|\le |T_A(x_j)|$. For $\forall D_i (1\le i\le m)$, the tolerance information granules of object x_j with respect to A and B can be rewritten as $T_A(x_j)=(T_A(x_j)\cap D_i)\cup(T_A(x_j)\cap(U-D_i))$ and $T_B(x_j)=(T_B(x_j)\cap D_i)\cup(T_B(x_j)\cap(U-D_i))$, respectively. Obviously, it holds that $|T_B(x_j)\cap D_i|\le |T_A(x_j)\cap D_i|$ and $|T_B(x_j)\cap(U-D_i)|\le |T_A(x_j)\cap(U-D_i)|$. Consequently, it follows from the monotonicity of the function $f(x,y)=-x\log(\frac{x}{x+y})$ [37], we have $-\frac{|T_B(x_j)\cap D_i|}{|U|}\log\frac{|T_B(x_j)\cap D_i|}{|T_B(x_j)|}\le -\frac{|T_A(x_j)\cap D_i|}{|U|}\log\frac{|T_A(x_j)\cap D_i|}{|T_A(x_j)|}$. Therefore, one can obtain that $H(D|B)=-\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{|T_B(x_j)\cap D_i|}{|U|}\log\frac{|T_B(x_j)\cap D_i|}{|T_B(x_j)|}\le H(D|A)=-\sum_{i=1}^{n}\sum_{j=}1^{m}\frac{|T_A(x_j)\cap D_i|}{|U|}\log\frac{|T_A(x_j)\cap D_i|}{|T_A(x_j)|}$. According to Definition 3, it holds that $I(D;A)\le I(D;B)$. This completes the proof.□

Theorem 1 states that the mutual information is monotone with respect to the set inclusion of features, i.e., the mutual information increases when the available features increase. That is to say, using more features gives more information about the output class. Therefore, the validity theorem guarantees that the mutual information can be used as a reasonable feature measure in incomplete decision systems.

4.2. Mutual information-based feature selection algorithm

In this subsection, a new feature evaluation function is proposed to overcome the drawbacks of existing evaluation functions for measuring candidate features. In addition, to speed up the feature selection process, the selection of candidate features can be implemented in a dwindling object set. On this basis, a mutual information-based feature selection algorithm is developed from incomplete data.

Designing a feature evaluation function to measure the quality of candidate features is one key issue in the feature selection process. However, most of the existing evaluation functions recognize the best candidate feature that only has the largest dependency on the target class [8,13,24], i.e., maximize $I(D;c)$, $\forall c\in C$, but not take into consideration the redundancy between the selected features. In fact, when two selected features highly depend on each other, the respective dependency on the target class would not change much if one of them is removed. In what follows, an illustrative example is given to describe some drawbacks of existing evaluation functions.

Example 1. Given an incomplete decision system $IDS=\langle U, A=C\cup D, V, f\rangle$, where $C=\{c_1, c_2, c_3, c_4\}$. Computed by the existing evaluation functions, the descending order of four candidate features is listed as follows:$I(D;c_2)>I(D;c_4)>I(D;c_3)>I(D;c_1)$. In the feature selection algorithm with greedy forward search, the feature subset containing features c_2, c_4andc_3 is selected. In fact, the subset containing features c_2 and c_3 is the optimal selection result. It is easy to verify that $I(D;c_2)=I(D;\{c_2,c_4\})$ by direct computation. The possible reasons for this computation include two aspects. One is that the feature c_4 is redundant. In fact, redundant features may affect the classification performance, but most existing evaluation functions do not take into consideration this aspect. The other is that the two features c_2 and c_4 depend on each other highly. If the feature c_4 is removed, it does not affect the dependency on the target class. However, most of the existing evaluation functions do not take into account the independency of selected features.

To overcome the above drawbacks, a new evaluation function for measuring candidate features is proposed as follows. The new evaluation function combines the maximal relevance and the minimal redundancy together by making use of the merits of the mRMR criterion [6,25].

Definition 5. Given an incomplete decision system $IDS=\langle U, A=C\cup D, V, f\rangle$, suppose $B\subseteq C$ is the selected feature subset, $a\in C-B$ is a candidate feature, the evaluation function of the candidate feature a is defined by

$$e(a)=e_1(a)-\frac{1}{|B|}e_2(a),\quad \text{where } e_1(a)=I(D;a) \text{ and } e_2(a)=I(B;a).$$

From this definition, the candidate feature that has high relevance with the target classes and low redundancy will be selected in the feature selection process. In this way, a reasonable subset of features can be selected.

Furthermore, there exist some strategies for speeding up the mutual information-based feature selection algorithm. To quicken the feature selection process, the selection of candidate features

can be implemented in a dwindling object set. Given an incomplete decision system $IDS=\langle U, A=C\cup D, V, f\rangle$, suppose $B\subseteq C$ is the subset of selected features, the objects in U can be classified into the following two parts: (1) the objects D_M that have been recognized with respect to B, and (2) the remaining objects $D_{\overline{M}}$ that cannot be recognized, where $D_M\cap D_{\overline{M}}=\varnothing and D_M\cup D_{\overline{M}}=U$. With regard to the mutual information, we have the following theorem.

Theorem 2. *Let $IDS=\langle U, A=C\cup D, V, f\rangle$ be an incomplete decision system, $U=\{x_1, x_2, \ldots, x_n\}$, $B\subseteq C$ is the selected feature subset, the classification induced by $BisU/TR(B)=\{T_B(x_1), T_B(x_2), \ldots, T_B(x_n)\}$, for the objects D_M recognized by B, if $a\in C-B$ is a candidate feature, then $I(D_M; B)=I(D_M; B\cup\{a\})$.*

Proof. Suppose the object set in the incomplete decision system be $U=\{x_1, x_2, \ldots, x_n\}$, the classification induced by $BisU/TR(B)=\{T_B(x_1), T_B(x_2), \ldots, T_B(x_n)\}$. For the objects D_M recognized by B, it follows from the definition of the mutual information that $I(D_M; B)=\sum_{j=1}^{n}\frac{|T_B(x_j)\cap D_M|}{|U|}\log\frac{|T_B(x_j)\cap D_M|}{|T_B(x_j)|}-\frac{|D_M|}{|U|}\log\frac{|D_M|}{|U|}$ and $I(D_M; B\cup\{a\})=\sum_{j=1}^{n}\frac{|T_{B\cup\{a\}}(x_j)\cap D_M|}{|U|}\log\frac{|T_{B\cup\{a\}}(x_j)\cap D_M|}{|T_{B\cup\{a\}}(x_j)|}-\frac{|D_M|}{|U|}\log\frac{|D_M|}{|U|}$. Since $B\subseteq B\cup\{a\}$, it holds that $T_{B\cup\{a\}}(x_j)\subseteq T_B(x_j)$. In addition, there are $T_B(x_j)=(T_B(x_j)\cap D_M)\cup(T_B(x_j)\cap(U-D_M))$, $T_{B\cup\{a\}}(x_j)=(T_{B\cup\{a\}}(x_j)\cap D_M)\cup(T_{B\cup\{a\}}(x_j)\cap(U-D_M))$, and B is the selected feature subset, thus it holds that $|T_B(x_j)\cap D_M|=|T_{B\cup\{a\}}(x_j)\cap D_M|$. According to the definition of the mutual information, one can obtain that $I(D_M; B)=I(D_M; B\cup\{a\})$. This completes the proof.□

Theorem 2 shows that if the objects have been recognized by the selected feature subset B, then any candidate feature a needs not to be worked on the objects. This indicates that the recognized objects D_M can be deleted from the whole object set U, and the objects gets much fewer as the selection of features goes on. Therefore, the selection of candidate features is in a dwindling object set. Based on this observation, we can mark the objects that have been recognized by the selected feature subset, and omit them in computing the mutual information. For example, we just need to analyze 45% of the objects to compute the mutual information if 55% of the objects that have been recognized by the selected feature subset. With this strategy, the computational overhead of the feature selection is greatly reduced. In particular, the effect is obvious for the incomplete decision systems where the size of object set is larger than that of the feature set, i.e., $|U|\gg|C|$.

On this basis, we will develop the mutual information-based feature selection algorithm with greedy forward search strategy from incomplete data as follows. The feature selection outline is given by Algorithm MIFS.

Algorithm 1. Mutual **i**nformation-based **f**eature **s**election algorithm from incomplete data (**MIFS**)

Input: An incomplete decision system $IDS=\langle U, A=C\cup D, V, f\rangle$, where $C=\{c_1, c_2, \ldots, c_m\}$;
Output: A feature subset Red
Begin
1. Initialize $Red=\varnothing; D_M=\varnothing; D_{\overline{M}}=U$;// The recognized objects D_M, the unrecognized objects $D_{\overline{M}}$
2. **For** $\forall c_i\in C(1\le i\le m)$, compute $sig(c_i, C, D)$;
3. *if* $sig(c_i, C, D)>0$, then $Red=Red\cup\{c_i\}$;
4. obtain the recognized objects D_M from $D_{\overline{M}}$ induced by c_i;
5. let $D_{\overline{M}}=D_{\overline{M}}-D_M$;
6. **End for**
7. **While** $D_{\overline{M}}\neq\varnothing$ do
8. compute $e(c)$ for all $c\in C-Red$;
9. choose the feature c_k that maximizes $e(c)$;
10. let $Red=Red\cup\{c_k\}; C=C-\{c_k\}$;
11. obtain the recognized objects D_M from $D_{\overline{M}}$ induced by c_k;
12. let $D_{\overline{M}}=D_{\overline{M}}-D_M$;
13. **End while**
14. **For** each $b\in Red$ **do**
15. compute $sig(b, Red, D)$;
16. if $sig(b, Red, D)=0$, then $Red=Red-\{b\}$;
17 **End for**
18 Return *Red*.

Complexity analysis of Algorithm MIFS: The following is the time complexity analysis of the algorithm. Here are some explanations first. Qian et al. [9] gave a fast algorithm for computing the tolerance classes with time complexity being $O(|C|^2|U|)$. Thus for large-scale incomplete data (where $|C|\ll|U|$), the time complexity of computing the mutual information is $O(|C|^2|U|+|U|+\sum_{i=1}^{n}|T_C(x_i)|\cdot\sum_{j=1}^{m}|D_j|)\approx k\cdot|U|^2$ (the specific introduction of $m, n, T_C(x_i)$, and D_j is shown in Definition 2), where k is the positive integer, $k\ll|U|$. Note that $\sum_{i=1}^{n}|T_C(x_i)|$ is approximately equal to $k\cdot|U|$, the main reason can be explained as follows. Since the worst case is $\sum_{i=1}^{n}|T_C(x_i)|=|U|\cdot|U|$, the feature values of the whole object set are all missing in this case. For such incomplete decision system, it is meaningless in real-world applications. Therefore, $\sum_{i=1}^{n}|T_C(x_i)|$ is approximately equal to $k\cdot|U|$.

Algorithm MIFS contains three main steps. At first, Steps 2–6 are to select the indispensable features from the whole feature set, and the objects recognized by the subset of selected features are deleted. The time complexity of Steps 2–6 is $O(k\cdot|C||U|^2)$. Then, Steps 7–13 are to add the current best feature c_k from remaining features to the selected feature subset by the proposed evaluation function $e(c)$, until the unrecognized objects is empty. In addition, the objects recognized by the subset of selected features are deleted in each loop. The time complexity of Steps 7–13 is

Table 1
Data sets description.

Data sets	Objects	Features	Classes
Cleveland	303	13	5
Breast cancer	699	10	2
Mfeat-factors	2000	216	10
Optdigits	3823	64	3
Satimage	6435	36	6
Mushroom	8124	22	2
Final-general	10,104	71	5
Gisette	13,500	5000	5
Semeion	1593	256	10
Isolet	6238	617	26

Table 2
Results of computational time.

Data sets	MIFS	CFS	IEFS	IGFS	mRMR	ReliefF
Cleveland	1.643	1.670	1.812	1.904	1.925	2.106
Breast cancer	4.291	6.514	8.033	8.725	9.130	9.958
Mfeat-factors	27.05	59.42	85.10	97.53	106.67	124.05
Optdigits	16.42	47.60	70.22	90.47	82.95	113.18
Satimage	41.78	96.91	110.59	143.60	129.47	162.65
Mushroom	35.94	74.05	92.38	108.71	115.23	140.79
Final-general	102.57	264.36	317.11	390.52	364.15	525.38
Gisette	218.39	425.27	503.94	635.16	587.04	831.57
Semeion	33.16	82.50	107.23	151.95	130.72	193.06
Isolet	51.02	106.79	148.62	179.21	125.83	204.17

$O(\sum_{i=1}^{|C|}(|C|-i+1)|U-D_M^i|\cdot max_{c\in C-Red, x\in D_M^i}|T_c(x)|)$, where D_M^i is the set of objects that have been recognized by the selected feature in the i-th loop, and D_M^i is the set of unrecognized objects in the i-th loop. Finally, Steps 14–17 are to delete the redundant features from the selected feature subset *Red* to guarantee the selection result *Red* without redundancy. The time complexity of Steps 14–17 is $O(k.|Red||U|^2)$ at most.

5. Experimental analysis

In this section, we will illustrate the efficiency and effectiveness of the proposed feature selection algorithm through experimental analysis. We have downloaded ten real-world data sets from UCI Repository of Machine Learning databases [55], which are described in Table 1. All the experiments are carried out on a PC with Windows7, Intel(R) Core(TM) Duo CPU2.93 GHz and 4 GB

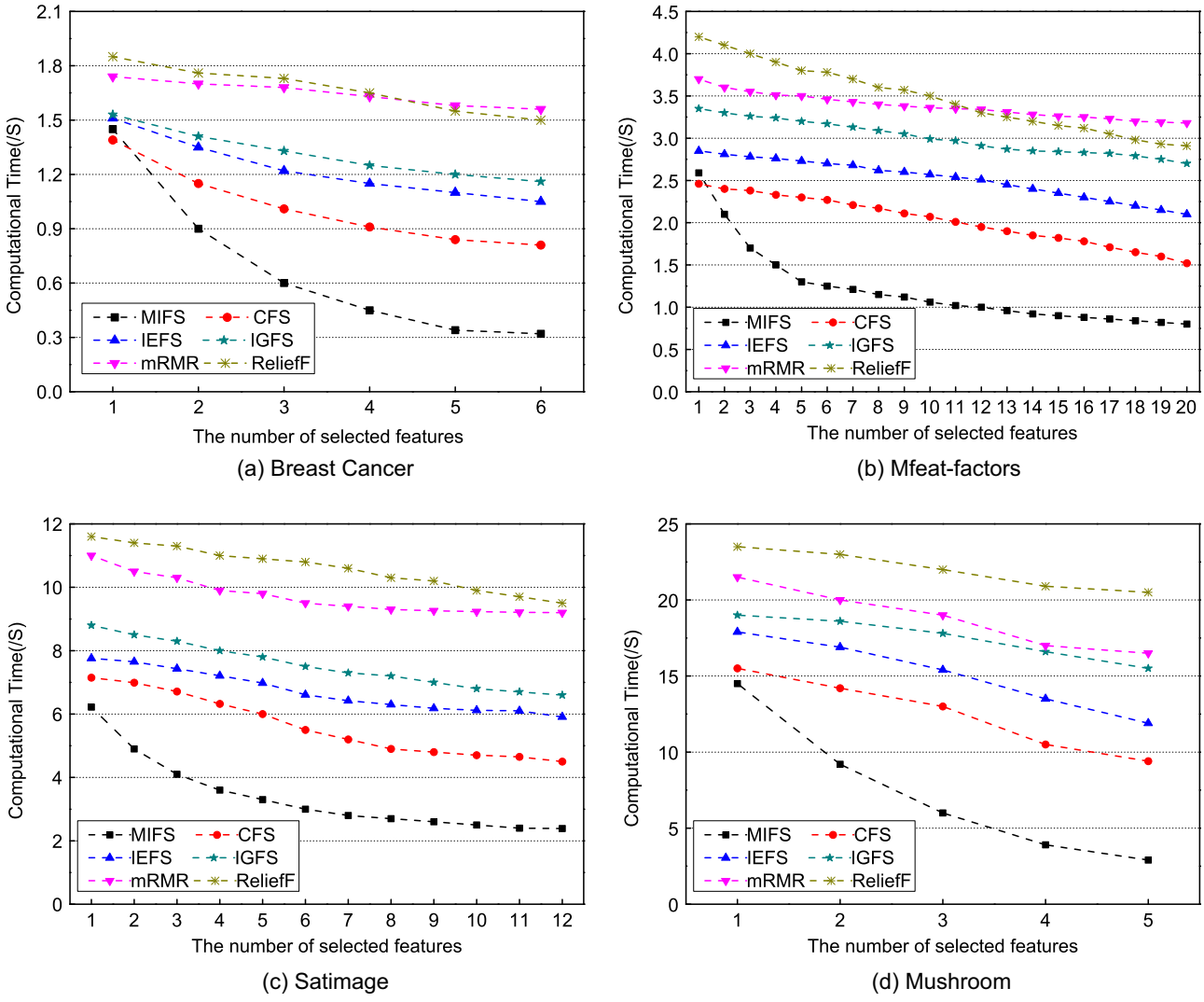

Fig. 1. Variation of computational time versus the order of selected features.

Table 3
Classification performance results by C4.5 classifier.

Data sets	UnSelect	MIFS	CFS	IEFS	IGFS	mRMR	ReliefF
Cleveland	51.64 ± 1.27	**57.52 ± 1.09**	55.21 ± 1.48	55.17 ± 1.12	56.09 ± 1.03	54.78 ± 2.32	53.91 ± 1.84
Breast cancer	83.15 ± 1.81	**91.01 ± 1.33**	89.42 ± 1.50	85.51 ± 1.74	86.72 ± 1.45	90.42 ± 1.57	82.63 ± 2.05
Mfeat-factors	75.29 ± 2.13	82.94 ± 1.85	**83.10 ± 1.77**	82.23 ± 1.92	80.51 ± 2.16	82.58 ± 1.90	79.74 ± 2.38
Optdigits	93.46 ± 0.85	96.27 ± 0.52	**96.73 ± 0.66**	95.70 ± 0.80	93.14 ± 0.78	95.13 ± 0.81	94.50 ± 0.93
Satimage	74.08 ± 1.97	**85.25 ± 1.64**	83.58 ± 1.37	84.16 ± 1.43	84.20 ± 2.23	83.71 ± 1.59	80.65 ± 2.17
Mushroom	96.93 ± 0.34	**100.0 ± 0.00**	**100.0 ± 0.00**	**100.0 ± 0.00**	**100.0 ± 0.00**	**100.0 ± 0.00**	98.71 ± 0.34
Final-general	71.52 ± 1.58	**76.38 ± 1.47**	75.16 ± 1.89	73.35 ± 2.10	73.67 ± 1.94	74.52 ± 2.46	72.60 ± 2.80
Gisette	50.31 ± 1.62	**69.70 ± 1.28**	68.25 ± 1.15	66.92 ± 1.67	68.58 ± 1.39	65.33 ± 1.96	61.94 ± 2.25
Semeion	88.20 ± 1.75	95.12 ± 0.94	92.68 ± 1.26	94.53 ± 1.18	91.07 ± 1.54	**95.56 ± 1.28**	89.32 ± 1.47
Isolet	69.55 ± 3.06	**82.47 ± 1.59**	80.29 ± 2.40	78.02 ± 2.51	80.51 ± 2.83	81.62 ± 1.75	75.41 ± 2.68
Ave	75.41	83.67	82.44	81.56	81.45	82.31	78.94

memory. Algorithms are coded in C++ and the software being used is Microsoft Visual 2008.

The data sets shown in Table 1 are used for evaluation purposes. For the complete data sets, we randomly take away 10% known features values of the original data set to create incomplete data sets. To diminish the effect of the randomness in classification performance, the missing feature values are taken from each feature equally from each complete data set, which can keep the data distribution. Other data sets that have missing values in the original version keep unchanged. In addition, for the data sets with numerical features, we use the data tool Rosetta (http://www.lcb.uu.se/tools/rosetta/index.php) to discretize them.

In the experiments, five representative feature selection algorithms are used to make comparisons with the proposed algorithm, namely, **c**onsistency-based approach **f**eature **s**election algorithm (CFS) [1], **i**nformation **e**ntropy-based approach **f**eature **s**election algorithm (IEFS) [3], **i**nformation **g**ain-based approach **f**eature **s**election algorithm (IGFS) [19], mRMR [6] and ReliefF [50]. Since the number of features selected by the feature selection algorithms is different, for the sake of impartiality, we choose the same quantity of features and the selected features are arranged in a descending order according to their priorities.

In what follows, we present the comparative performance analysis of the six feature selection algorithms from two aspects: computational time and classification accuracy. In the experiments, two popular classifiers, namely C4.5 and Naive Bayes are chosen to test the prediction accuracies of the feature subsets selected by different feature selection algorithms.

5.1. Computational time

Table 2 records the computational time of six feature selection algorithms for selecting the feature subsets. The computational time is expressed as seconds.

All the results reported in Table 2 establish the fact that the computational time of six algorithms increases as the size of data sets increases. However, the proposed algorithm MIFS selects the feature subsets with less time than other feature selection algorithms, irrespective of the data sets. For example, for the data set Satimage, CFS, IEFS, IGFS, mRMR and ReliefF take 96.91 s, 110.59 s, 143.60 s, 129.47 s, and 162.65 s to select the feature subset, respectively. In contrast, MIFS takes about 41.78 s to find the feature subset. The time needed by MIFS is much shorter than other methods. In addition, take the data set Gisette as an example, MIFS needs only 218.39 s to find the feature subset, while all of the algorithms CFS, IEFS, IGFS, mRMR and ReliefF use more than twice the time than that of MIFS. The similar behaviors also hold for other data sets. The advantage of MIFS over other methods is clear, particularly for large-scale data sets. This is expected since the proposed algorithm involves a dwindling object set.

To further investigate the efficiency of the proposed feature selection algorithm, we select four data sets Breast Cancer, Mfeat-factors, Satimage, and Mushroom to conduct an experimental comparison among six feature selection algorithms. More detailed change trendline of each algorithm is displayed in Fig. 1. In the following figure, the *x*-coordinate pertains to the number of selected features, while *y*-coordinate concerns the computational time.

From Fig. 1, the differences of the six algorithms on the data sets are not distinctly different at the beginning, and then the curve of MIFS drops distinctly. That is to say, the time cost for MIFS to select features is much less than other algorithms. Obviously, with the increase of the selected features in all data sets, the computational saving caused by the MIFS algorithm becomes increasingly significant. Take the Mfeat-factors data set for example, MIFS takes about 0.9 s to select the twelfth feature, while CFS, IEFS, IGFS, mRMR and ReliefF take about 1.95 s, 2.51 s, 2.91 s, 3.36 s, and 3.28 s at the same situation, respectively. The main reason is that the number of the unrecognized objects becomes less and less in MIFS, since the marked objects are deleted from the universe gradually in the feature selection process. Compared with other algorithms, the selection of candidate features in MIFS is implemented in a dwindling object set with less number. However, the computational time of other five feature selection algorithms slowly decrease, and the decrease slows down until it completely stops. This phenomenon happens attributed to the reason that the number of objects in the five algorithms CFS, IEFS, IGFS, mRMR and ReliefF almost keeps unchanged when selecting candidate features. The similar behaviors also hold for other three data sets. The experimental results indicate that MIFS is more efficient than other five algorithms for feature subset selection from incomplete data.

5.2. Classification accuracy

In what follows, the classification performance of the proposed algorithm, along with a comparison with other five algorithms, is demonstrated on the data sets using the predictive accuracy of the two classifiers: C4.5 and Naive Bayes.

In the experiments, for the purpose of illustrating the quality of subsets discovered by the six feature selection algorithms, we follow a 10-fold cross validation strategy to evaluate different algorithms. Each data set is divided into two disjoint parts by random sampling: one for training and the other for test. In each fold, we use the feature selection algorithms to reduce the training set. After selecting the feature subsets, the classifiers C4.5 and Naive Bayes are employed to extract the rules from the reduced training set. The test set is used to test the classification performance of the extracted rules. On the test set, the presented classification results are obtained. The numbers shown in Tables 3 and 4 are the average classification accuracy over the

Table 4
Classification performance results by Naive Bayes classifier.

Data sets	UnSelect	MIFS	CFS	IEFS	IGFS	mRMR	ReliefF
Cleveland	55.02 ± 0.94	**59.22 ± 0.85**	56.81 ± 1.17	58.65 ± 0.99	58.92 ± 1.20	56.17 ± 1.63	51.84 ± 1.25
Breast cancer	96.41 ± 0.51	**97.11 ± 0.46**	96.30 ± 0.58	94.03 ± 0.72	94.47 ± 0.93	97.08 ± 0.82	92.71 ± 1.01
Mfeat-factors	78.57 ± 1.65	**84.15 ± 1.30**	84.03 ± 1.35	82.19 ± 1.38	83.10 ± 1.72	83.44 ± 1.95	80.53 ± 1.47
Optdigits	90.16 ± 0.82	98.43 ± 0.61	96.72 ± 0.64	96.13 ± 0.57	98.52 ± 0.45	**98.90 ± 0.77**	93.25 ± 0.59
Satimage	72.35 ± 2.06	**83.50 ± 1.72**	82.35 ± 1.90	80.27 ± 2.15	77.61 ± 2.33	82.18 ± 2.40	78.62 ± 2.93
Mushroom	98.12 ± 0.32	**100.0 ± 0.00**	**100.0 ± 0.00**	**100.0 ± 0.00**	**100.0 ± 0.00**	**100.0 ± 0.00**	99.18 ± 0.21
Final-general	67.68 ± 1.79	71.92 ± 1.54	71.38 ± 1.70	**72.25 ± 1.64**	69.18 ± 1.61	72.03 ± 1.87	68.94 ± 2.36
Gisette	53.40 ± 1.30	**76.08 ± 1.22**	72.55 ± 1.42	73.41 ± 1.35	75.35 ± 1.19	75.46 ± 1.59	72.80 ± 1.74
Semeion	90.95 ± 0.83	97.65 ± 0.51	**98.04 ± 0.69**	95.38 ± 0.73	87.55 ± 1.04	95.28 ± 0.96	92.52 ± 1.26
Isolet	73.58 ± 3.91	**85.24 ± 2.47**	79.63 ± 2.85	82.71 ± 2.32	77.94 ± 3.53	80.45 ± 2.87	74.69 ± 3.38
Ave	77.62	85.33	83.78	83.50	82.26	84.10	80.51

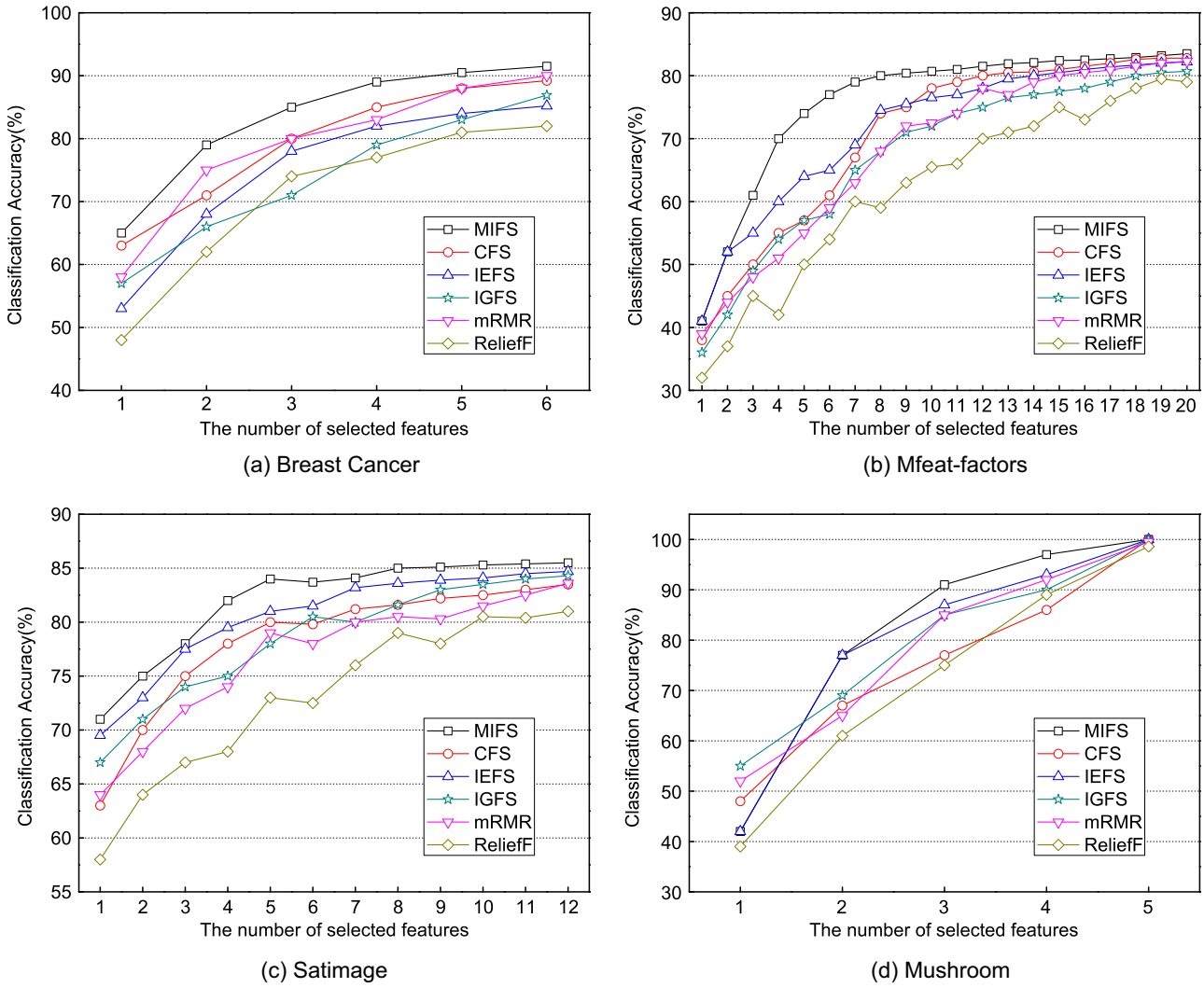

Fig. 2. Variation of classification accuracies by C4.5 with the number of selected features.

10-fold cross validation and the standard deviation of the accuracy obtained in those runs. The average classification accuracies are expressed in percentage. The "Ave" row records the average value of classification accuracy induced by each algorithm on ten data sets. The boldfaced values are the highest ones.

From the results reported in Tables 3 and 4, it is seen that through six feature selection algorithms, there is no remarkably large deterioration in the classification performance when compared with the results from the unselected data set. The results shows that all the feature selection algorithms are effective in retaining the classification accuracies through the removal of noisy, irrelevant, or redundant features. Also, it is shown that the proposed feature selection algorithm is more powerful than other feature selection algorithms in the comparative study. Specifically, as for the C4.5 classifier, MIFS is significantly better than the algorithms CFS, IEFS, IGFS, mRMR and ReliefF in seven, nine, nine, eight, and ten cases, respectively, out of ten cases in total. In addition, from the average classification accuracies presented in the last row of Table 3, it can be seen that the average accuracy value for all data sets is equal to 82.44% for CFS, 81.56% for IEFS, 81.45% for IGFS, 82.31% for mRMR, and 78.94% for ReliefF, compared with 83.67% for MIFS. From this, we can observe that the proposed algorithm is clearly superior to others on most of the data sets. The better performance of the proposed algorithm is achieved due to the fact that the correlation effect of the selected features in the MIFS algorithm can lead to good classification performance. A subset of features that are independent of each other would be less optimal. With the experimental results presented in Table 4 by the Naive Bayes classifier, we can see that the proposed algorithm MIFS shows similar patterns to that of C4.5 classifier. There is also a significant improvement in accuracy value in most of the data sets when comparing with other algorithms. This indicates that the proposed algorithm can achieve its best performance, irrespective of the classifier used.

Figs. 2 and 3 show the variation of classification accuracies with the number of selected features obtained for data sets Breast Cancer, Mfeat-factors, Satimage and Mushroom. The feature subsets are selected by the six feature selection algorithms MIFS, CFS, IEFS, IGFS, mRMR and ReliefF, respectively. In the following figures, the x-coordinate pertains to the number of selected features, while y-coordinate concerns the variation of classification accuracy.

From Fig. 2, it can be seen that with different number of selected features, MIFS always gives the classification performance that is better than or comparable to that of other algorithms. As for data set Satimage, the difference of the accuracy values of the six algorithms is not obvious at the beginning of the selection process, and this phenomenon occurs for the Breast Cancer, Satimage, and Mushroom data sets. And then the classification accuracy of MIFS is better than or comparable to that of other algorithms. This result indicates that the mutual information criterion used for selecting candidate features in the feature selection process provides high classification performance. The success of the mutual

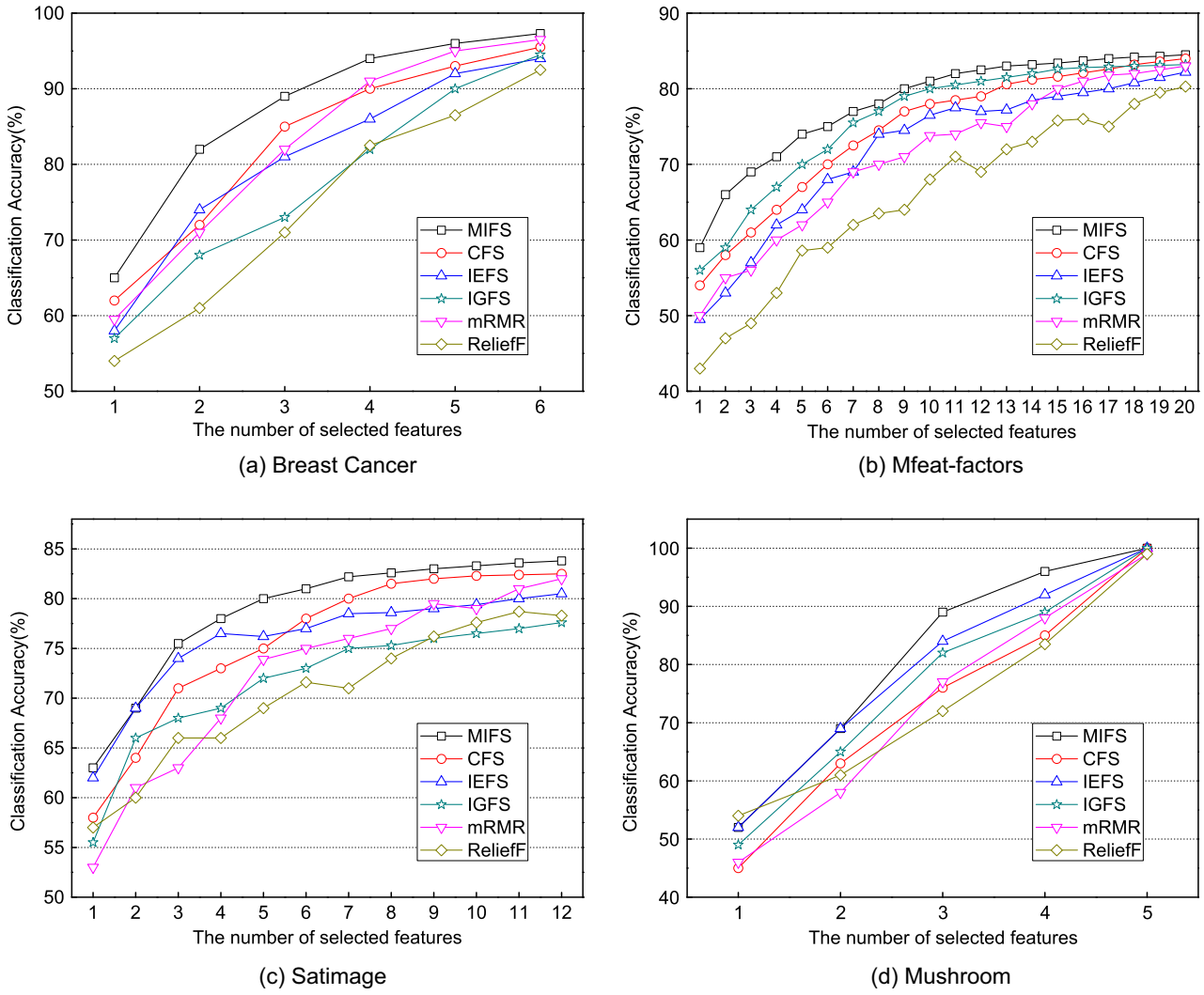

Fig. 3. Variation of classification accuracies by Naive Bayes with the number of selected features.

Table 5
Classification accuracy comparison of C4.5 with different discretization methods.

Data sets	MDL	FCM	Eq-Width	Eq-Freq
Cleveland	57.52 ± 1.09	56.83 ± 2.30	54.96 ± 2.14	56.07 ± 1.95
Breast cancer	91.01 ± 1.33	90.42 ± 1.25	91.61 ± 2.60	93.18 ± 3.07
Mfeat-factors	82.94 ± 1.85	83.50 ± 2.46	82.29 ± 1.75	82.29 ± 1.75
Optdigits	96.27 ± 0.52	95.91 ± 1.13	94.87 ± 1.09	95.42 ± 2.23
Satimage	85.25 ± 1.64	83.72 ± 1.72	85.14 ± 2.36	85.36 ± 1.89
Mushroom	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00
Final-general	76.38 ± 1.47	74.60 ± 2.28	75.35 ± 2.61	76.54 ± 1.70
Gisette	69.70 ± 1.28	68.54 ± 3.69	70.26 ± 1.87	71.63 ± 2.51
Semeion	95.12 ± 0.94	95.12 ± 0.94	95.12 ± 0.94	95.12 ± 0.94
Isolet	82.47 ± .59	80.95 ± 2.57	81.43 ± 2.35	81.75 ± 1.46
Ave	83.67	82.96	83.10	83.74

Table 6
Classification accuracy comparison of Naive Bayes with different discretization methods.

Data sets	MDL	FCM	Eq-Width	Eq-Freq
Cleveland	59.22 ± 0.85	59.04 ± 1.37	60.52 ± 2.10	60.68 ± 1.49
Breast cancer	97.11 ± 0.46	93.18 ± 0.51	96.30 ± 1.03	97.45 ± 0.81
Mfeat-factors	84.15 ± 1.30	81.25 ± 1.90	83.61 ± 1.74	83.61 ± 1.74
Optdigits	98.43 ± 0.61	96.79 ± 1.24	97.40 ± 0.86	98.12 ± 1.13
Satimage	83.50 ± 1.72	84.32 ± 2.35	82.94 ± 2.57	83.06 ± 1.91
Mushroom	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00
Final-general	71.92 ± 1.54	69.77 ± 2.06	68.95 ± 1.82	71.23 ± 1.68
Gisette	76.08 ± 1.22	76.08 ± 1.22	76.08 ± 1.22	76.08 ± 1.22
Semeion	97.65 ± 0.51	96.41 ± 1.15	96.73 ± 0.90	97.55 ± 0.87
Isolet	85.24 ± 2.47	84.63 ± 2.81	85.11 ± 2.04	85.07 ± 2.23
Ave	85.33	84.15	84.76	85.28

information-based feature selection algorithm depends critically on much information about the output class contained in the selected features. Occasionally, the accuracy values of other five algorithms decrease a little when compared with MIFS. This phenomenon happens attributed to some features selected by other five algorithms are superfluous for classification, which may deteriorate the classification performance. Their impact is negative, and they should be eliminated from the selected feature subsets. But for MIFS, there is a redundancy-removing step. Some redundant feature can be deleted from the selection results by this step. As to other three data sets, one may observe the same situation. A similar behavior happens to Fig. 3. The results indicate that MIFS has good classification performance in feature selection from incomplete data.

5.3. Sensitivity analysis

In what follows, we further perform the experiments to analyze the sensitivity analysis of the feature selection results to different

discretization methods. In this experiment, we use the minimum descriptive length (MDL) method [51] to discretize the numerical features. For the experimental study, other three discretization techniques are made a comparison with the MDL method: fuzzy *c*-means (FCM) [52], Equal frequency (Eq-Freq) [53], and Equal Width (Eq-Width) [54]. For the Eq-Freq and Eq-Width discretization methods, the number of intervals is set to 10. Then we conduct the proposed algorithm on the discretized data sets.

Tables 5 and 6 present the comparison of classification performance of C4.5 and Naive Bayes with different discretization methods, respectively. Average classification accuracies of all data sets are shown at the last row, which indicate how the discretization methods affect predictive accuracy.

From Tables 5 and 6, it is shown that most discretization methods do not significantly decrease classification accuracies. In Table 5, as the discretized data with the MDL and Eq-Freq methods, we can see that the feature subsets selected can produce slightly higher classification accuracies in most of data sets. Similarly among the data sets, FCM and Eq-Width methods show slightly lower classification accuracies in most data sets. The MDL and Eq-Freq methods perform better than FCM and Eq-Width methods. As to two data sets Mushroom and Semeion, four discretization methods have the same classification results, the reason is that two data sets have no numerical features. It can be noted from Table 5 that the average accuracy value obtained for MDL is 83.67%, while the values obtained for FCM, Eq-Width and Eq-Freq are 82.96%, 83.10% and 83.74%, respectively. From these results, we know that the average classification accuracies for all the three discretization methods change little. The discretization methods behave similarly for most of the data sets.

Corresponding to the Naive Bayes shown in Table 6, the similar behaviors of different discretization methods hold. Except for Mushroom and Semeion data sets, it can be seen that MDL method has higher classification accuracies in five out of eight data sets, while the Eq-Freq method has higher classification accuracies only in two data sets. The similar case occurs to FCM and Eq-Width based discretization method. Therefore, we could envision that MDL is relatively superior to other discretization methods at most case. We can also see that, as for data set Mfeat-factors, Eq-Width and Eq-Freq methods have the same results, the main reason is that the selected feature subsets of two discretization methods are the same. In addition, from the last low of Table 6, we can see that the average accuracy value of MDL is 85.33%, while the values of FCM, Eq-Width and Eq-Freq methods are 84.15%, 84.76% and 85.28%, respectively. No one discretization method can ensure an obvious advantage in most cases. As the experimental results shown in Tables 5 and 6, the results demonstrate that our proposed algorithm is not sensitive to the feature discretion methods at most cases.

From the aforementioned experimental results, the following conclusion can be drawn that the proposed feature selection algorithm can achieve good classification accuracy with a reasonably compact set of features, which provides an efficient solution for feature selection in incomplete data environment.

6. Conclusions

The contribution of the paper lies in developing an efficient mutual information-based feature selection algorithm from incomplete data, which integrates the information theory and rough sets. In the process of feature selection, the proposed evaluation function can select candidate features that have high relevance to the class and low redundancy among the selected features, such that the redundancy is eliminated. In addition, the selection of candidate features is implemented in a dwindling object set, such that the computational efforts are greatly reduced. Numerical experiments of the proposed method, in comparison with some existing methods, are provided for different data sets. The experimental results show that the proposed algorithm is effective for feature selection from incomplete data. In our further research, we will extend the proposed approach to handle hybrid data in which both the categorical data and numeric data coexist. In addition, how to further improve the feature selection algorithm with a parallel strategy from incomplete big data is another interesting topic.

Acknowledgments

We would like to thank the anonymous referees and editors for their valuable comments and suggestions. This work is supported by the Natural Science Foundation of Jiangxi Province (No. 20151BAB217009), and the Natural Science Foundation of China (Nos. 71461013 and 61462037).

References

[1] A. Arauzo-Azofra, J.M. Benitez, J.L. Castro, Consistency measures for feature selection, J. Intell. Inf. Syst. 30 (3) (2008) 273–292.
[2] M. Kolar, H. Liu, Feature selection in high-dimensional classification, in: Proceedings of the 30th International Conference on Machine Learning, 2013, pp. 329–337.
[3] L. Sun, J.C. Xu, Y. Tian, Feature selection using rough entropy-based uncertainty measures in incomplete decision systems, Knowl. Based Syst. 36 (2012) 206–216.
[4] Q.H. Hu, X.J. Che, L. Zhang, D.R. Yu, Feature evaluation and selection based on neighborhood soft margin, Neurocomputing 73 (10–12) (2010) 2114–2124.
[5] A.K. Jain, D. Zongker, Feature selection: evaluation, application and small sample performance, IEEE Trans. Pattern Anal. Mach. Intell. 19 (2) (1997) 153–158.
[6] 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.
[7] D. Huang, T.W.S. Chow, Effective feature selection scheme using mutual information, Neurocomputing 63 (2005) 325–343.
[8] 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.
[9] Y.H. Qian, J.Y. Liang, W. Pedrycz, C.Y. Dang, An efficient accelerator for attribute reduction from incomplete data in rough set framework, Pattern Recognit. 44 (2011) 1658–1670.
[10] D. Tian, X.J. Zeng, J. Keane, Core-generating approximate minimum entropy discretization for rough set feature selection in pattern classification, Int. J. Approx. Reason. 52 (2011) 863–880.
[11] Z. Pawlak, Rough Sets, Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, 1991.
[12] M. Kryszkiewicz, Rough set approach to incomplete information system, Inf. Sci. 112 (1998) 39–49.
[13] P.A. Estevez, M. Tesmer, C.A. Perez, J.M. Zurada, Normalized mutual information feature selection, IEEE Trans. Neural Netw. 20 (2) (2009) 189–201.
[14] F.F. Xu, D.Q. Miao, L. Wei, Fuzzy-rough attribute reduction via mutual information with an application to cancer classification, Comput. Math. Appl. 57 (2009) 1010–1017.
[15] A. Aussem, S.R.D. Morais, A conservative feature subset selection algorithm with missing data, Neurocomputing 73 (4–6) (2010) 585–590.
[16] J.H. Dai, W.T. Wang, Q. Xu, An uncertainty measure for incomplete decision tables and its applications, IEEE Trans. Cybern. 43 (4) (2013) 1277–1289.
[17] Q. Mao, I.W.H. Tsang, Optimizing performance measures for feature selection, in: IEEE 11th International Conference on Data Mining, 2011, pp. 1170–1175.
[18] N.M. Parthalain, Q. Shen, Exploring the boundary region of tolerance rough sets for feature selection, Pattern Recognit. 42 (2009) 655–667.
[19] 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.
[20] Q.B. Song, J.J. Ni, G.T. Wang, A fast clustering-based feature subset selection algorithm for high-dimensional data, IEEE Trans. Knowl. Data Eng. 25 (1) (2013) 1–14.
[21] B. Zhang, L. Zhang, Theory and application of problem solving, North-Holland, Amsterdam, 1992.
[22] Z. Zhu, Y.S. Ong, M. Dash, Wrapper-filter feature selection algorithm using a memetic framework, IEEE Trans. Syst. Man Cybern.: Part B 37 (1) (2007) 70–76.
[23] H.W. Liu, J.G. Sun, L. Liu, H.J. Zhang, Feature selection with dynamic mutual information, Pattern Recognit. 42 (2009) 1330–1339.
[24] J.B. Yang, C.J. Ong, An effective feature selection method via mutual information estimation, IEEE Trans. Syst. Man Cybern.—Part B: Cybern. 42 (6) (2012) 1550–1559.

[25] P. Maji, P. Garai, On fuzzy-rough attribute selection: criteria of max-dependency, max-relevance, min-redundancy, and max-significance, Appl. Soft Comput. 13 (9) (2013) 3968–3980.
[26] R.W. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognit. Lett. 24 (2003) 833–849.
[27] L. Yun, L.L. Bao, Feature selection based on loss-margin of nearest neighbor classification, Pattern Recognit. 42 (2009) 1914–1921.
[28] P. Hajek, Z. Hanikova, A set theory within fuzzy logic, In: Proceedings of 31st IEEE International Symposium on Multiple-Valued Logic, 2001, pp. 319–323.
[29] H. Liu, H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Kluwer Academic Publishers, Boston, 1998.
[30] P. Somol, P. Pudil, J. Kittler, Fast branch and bound algorithms for optimal feature selection, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2004) 900–912.
[31] L.J. Ke, Z.R. Feng, Z.G. Ren, An efficient ant colony optimization approach to attribute reduction in rough set theory, Pattern Recognit. Lett. 29 (2008) 1351–1357.
[32] J.Q. Gan, B.A.S. Hasan, C.S.L. Tsui, A filter-dominating hybrid sequential forward floating search methods for feature subset selection in high-dimensional space, Int. J. Mach. Learn. Cybern. 5 (3) (2014) 413–423.
[33] K.Z. Mao, Orthogonal forward selection and backward elimination algorithms for feature subset selection, IEEE Trans. Syst. Man Cyber.—Part B: Cybern. 34 (1) (2004) 629–634.
[34] Z. Meng, Z. Shi, A fast approach to attribute reduction in incomplete decision systems with tolerance relation-based rough sets, Inf. Sci. 179 (2009) 2774–2793.
[35] Y. Qian, J. Liang, D. Li, F. Wang, N. Ma, Approximation reduction in inconsistent incomplete decision tables, Knowl. Based Syst. 23 (2010) 427–433.
[36] L. Zadeh, Some reflections on soft computing, granular computing and their roles in the conception, design, and utilization of information/ intelligent systems, Soft Comput. 2 (1998) 23–25.
[37] J. Dai, W. Wang, Q. Xu, H. Tian, Uncertainty measurement for interval-valued decision systems based on extended conditional entropy, Knowl. Based Syst. 27 (2012) 443–450.
[38] A. Skowron, J. Stepaniuk, R. Swiniarski, Modeling rough granular computing based on approximation spaces, Inf. Sci. 184 (2012) 20–43.
[39] W.G. Zhou, C.G. Zhou, H. Zhu, G.X. Liu, X.Y. Chang, Feature selection for microarray data analysis using mutual information and rough set theory, IN: International Conference on Intelligent Computing, 2006, pp. 424–432.
[40] J.W. Grzymala-Busse, P.G. Clark, M. Kuehnhausen, Generalized probabilistic approximations of incomplete data, Int. J. Approx. Reason. 55 (1) (2014) 180–196.
[41] M. Sebban, R. Nock, A hybrid filter/wrapper approach of feature selection using information theory, Pattern Recognit. 35 (4) (2002) 835–846.
[42] W.H. Shu, W.B. Qian, A fast approach to attribute reduction from perspective of attribute measures in incomplete decision systems, Knowl. Based Syst. 72 (2014) 60–71.
[43] W.Z. Wu, M. Zhang, H.Z. Li, J.S. Mi, Knowledge reduction in random information systems via Dempster–Shafer theory of evidence, Inf. Sci. 174 (3–4) (2005) 143–164.
[44] V.G. Verdejo, M. Verleysen, J. Fleury, Information-theoretic feature selection for functional data classification, Neurocomputing 72 (16–18) (2009) 3580–3589.
[45] Q. He, Z.X. Xie, Q.H. Hu, C. Wu, Neighborhood based sample and feature selection for SVM classification learning, Neurocomputing 74 (10) (2011) 1585–1594.
[46] W. Pedrycz, A. Bargiela, An optimization of allocation of information granularity in the interpretation of data structures: toward granular fuzzy clustering, IEEE Trans. Syst. Man Cybern.-Part B: Cybern. 42 (3) (2012) 582–590.
[47] D. Francois, F. Rossi, V. Wertz, M. Verleysen, Resampling methods for parameter-free and robust feature selection with mutual information, Neurocomputing 70 (7–9) (2007) 1276–1288.
[48] Y.Y. Yao, Interpreting concept learning in cognitive informatics and granular computing,, IEEE Trans. Syst. Man Cybern.-Part B 39 (4) (2009) 855–866.
[49] M. Yang, P. Yang, A novel condensing tree structure for rough set feature selection, Neurocomputing 71 (4–6) (2008) 1092–1100.
[50] M.R. Sikonja, I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, Machine Learning 53(1-2) 53 (1–2) (2003) 23–69.
[51] U.M. Fayyad, K.B. Irani, Multi-interval discretization of continuous valued attributes for classification learning, in: Proceedings of the 13th International Joint Conference on Artificial Intelligence, 1993, pp. 1022–1027.
[52] Y. Yang, G.I. Webb, A comparative study of discretization methods for Naive Bayes classifiers, in: Proceedings of Pacific Rim Knowledge Acquisition Workshop, 2002, pp. 159–173.
[53] H. Liu, F. Hussain, C.L. Tan, M. Dash, Discretization: an enabling technique, Data Min. Knowl. Discov. 6 (2002) 393–423.
[54] M. Boulle, Khiops: A statistical discretization method of continuous attributes, Mach. Learn. 55 (2004) 53–69.
[55] UCI Machine Learning Repository, ⟨http://www.ics.uci.edu/mlearn/MLRepository.html⟩.

Wenbin Qian received the Ph.D. degree in computer science at University of Science and Technology Beijing, China, in 2014. He is currently a Lecturer with the School of Software of Jiangxi Agriculture University, China, and is also a Co-investigator at Beijing Key Laboratory of Knowledge Engineering for Materials Science. His current research interests include data mining, rough sets and computational intelligence.

Wenhao Shu received the Ph.D. degree in the School of Computer and Information Technology of Beijing Jiaotong University, China, in 2015. She is currently a Lecturer with the School of Information Engineering of East China, Jiaotong University, China, Her research interests include knowledge discovery, rough sets and machine learning.