J Intell Inf Syst (2013) 41:235–268
DOI 10.1007/s10844-013-0243-x

A novel feature selection method and its application

Bing Li · Tommy W. S. Chow · Di Huang

Received: 11 April 2012 / Revised: 8 March 2013 / Accepted: 11 March 2013 /
Published online: 4 April 2013

Abstract In this paper, a novel feature selection method based on rough sets and mutual information is proposed. The dependency of each feature guides the selection, and mutual information is employed to reduce the features which do not favor addition of dependency significantly. So the dependency of the subset found by our method reaches maximum with small number of features. Since our method evaluates both definitive relevance and uncertain relevance by a combined selection criterion of dependency and class-based distance metric, the feature subset is more relevant than other rough sets based methods. As a result, the subset is near optimal solution. In order to verify the contribution, eight different classification applications are employed. Our method is also employed on a real Alzheimer's disease dataset, and finds a feature subset where classification accuracy arrives at 81.3 %. Those present results verify the contribution of our method.

Keywords Alzheimer's disease · Class-based distance metric · Feature selection · Mutual information · Rough sets

1 Introduction

The size of dataset has been increasing dramatically, which usually incurs high computational costs, so handling huge dataset and large dimensional dataset have

B. Li (✉) · T. W. S. Chow
Department of Electronic Engineering, City University of Hong Kong,
83 Tat Chu Avenue, Kowloon, Hong Kong
e-mail: bingli5@student.cityu.edu.hk; lib675@gmail.com

D. Huang
Computational Biology Branch, National Center for Biotechnology Information,
National Library of Medicine, National Institutes of Health,
8600 Rockville Pike, Bethesda, MD 20894, USA

become a major problem (Kumar 1998; Aouiche and Darmont 2009). In order to deal with this problem, feature reduction and data reduction are practically important. Feature selection (FS), as a kind of feature reduction, can find a more relevant feature subset with labels. Since FS reduces the number of features, it can save cost of computational time and memory when dealing with high dimensional datasets. It is also useful to improve classification accuracy as a result of removing redundant and irrelevant features. According to different mechanisms of selection, FS methods fall into three catalogues: wrapper feature selection (Kuncheva and Jain 1999; Ron and John 1997), embedded method (Breiman et al. 1984; Quinlan 1993), and filter method (Xu et al. 2010; Hu et al. 2010; Dash and Liu 2003; Liu et al. 2009; Peng et al. 2005; Sotoca and Pla 2010). Since filter methods can support the balance between classification accuracy and computational cost, it is used most widely. Two kinds of filter methods based on mutual information (MI) or rough sets are introduced below, because the theories are also employed in our methods.

Since MI can estimate the certainty between two variables, it is used as a criterion to select features. The main idea of MI based methods finds a subset by maximizing MI between features and labels, and minimizing MI among features. On the basis of this idea, each method defines its own selection criterion. The basic methods, Mutual information based feature selection (MIFS) (Battiti 1994) and MIFS-U (Kwak and Choi 2002a), consider the certainty between labels and features. Feature selection by mutual information based on parzen window (PWMI) eliminates features with little information about class variable (Kwak and Choi 2002b). For optimal feature subset using MI (OFS-MI), salient features are selected by comparing the quadratic MI (Chow and Huang 2005; Huang and Chow 2005). Normalized MIFS (Estévez et al. 2009) is a fast and efficient method, due to its incremental nature. For MI methods, a serious problem is that the subset found by ‘best individual feature’ does not imply that it is the best subset of features (Cover and Thomas 1991). That is to say that the subset selected by MI cannot verify the combination of features due to large computational cost of high dimensional MI. In contrast, rough sets based feature selection delivers encouraging results in combination of features.

Rough set is a mathematical theory developed by Pawlak (1982, 1991) and Pawlak and Skowron (2007a, b, c). The process of standard feature selection employing rough sets can be viewed as reduct construction (Pawlak 1982). Comparing with other feature selection methods, reduct construction defines its own selection and termination criteria (Yao et al. 2008). A method of reduct construction does not stop until it finds a set where the combination of each feature with others favors the increasing of dependency. There are extensions about rough sets base FS methods. Samples in the Boundary Region are employed to define a distance metric as part of selection criterion (Parthaláin et al. 2007; Parthaláin and Shen 2009). The subsets selected by these methods include more information about labels. Heuristic functions guide the searching process of FS methods (Zhong et al. 2001; ElAlami 2009; Bae et al. 2010; Wang et al. 2007; Chen et al. 2011), which favor finding a more desirable output. Heuristic functions depend on internal information of data. In contract, external information based on semantics or constraints (Yao et al. 2006) is used to find a reduct with user preference. Hybrid fuzzy and rough method (Hu et al. 2007; Parthaláin et al. 2010; Cornelis et al. 2008, 2010a, b; Jensen et al. 2009; Cornelis and Jensen 2008) can provide flexible solutions by extending lower and upper approximations of rough sets to fuzzy sets. Finally, a feature selection algorithm for multiple classifiers can decrease the number of decision-relative feature subsets (Delimata and Suraj 2008). This method evaluates the feature subsets according to deterministic and inhibitory rules which consider the influence of an added new object on decision table (Moshkov et al. 2008, 2010; Delimata et al. 2008, 2009, 2010).

Rough sets based FS methods are filter methods, but the concept of rough sets is classification, which is similar with wrapper methods. That is to say that the subsets selected by rough sets methods favor the classification accuracy without dramatical improvement of computational cost. However, it is a NP-hard problem to find an optimal solution. In this paper, we propose a new algorithm to find a near optimal subset. Since dependency expresses the ability to discern feature values, it is set to guide the searching process. To avoid selecting the features with small addition of dependency, MI is employed to evaluate the similarities of features. As a result, the dependency of the subset found by our method reaches the maximum with small number of features. The dependency shows the definitive relevance of the samples in the Positive Region. But the relationship between the samples in the Boundary Region and labels is not considered by the dependency. A class-based distance metric (CDM), which is part of selection criterion, is defined to evaluate the benefit of the Boundary Region on classification accuracies.

Our proposed feature selection based on rough sets and mutual information (RSMI-FS) shows four major advantages. First, since the dependency of each single feature as heuristic information guides the searching process, the subset determined by RSMI-FS is optimal or near-optimal. This is confirmed by our results in Section 4. For each dataset, the classification accuracies of the subsets found by our method are larger or comparable comparing with other methods. Second, RSMI-FS can find more relevant features, since its selection criterion, which combines the dependency with CDM, evaluates the definitive and uncertain relevance between features and labels. Third, MI is used to reduce the redundant features. Fourth, the optimal subset is found without dramatical increasing of computational time cost, as quadratic MI is used to estimate the similarities of features, which can reduce the computational complexity of MI (Chow and Huang 2005; Huang and Chow 2005).

A real application dataset is also used to verify the effectiveness of our method. Estimated 5.4 million Americans have Alzheimer's disease (AD). Every 68 s, an additional person with AD is found now. By 2050, there will be one new case of AD every 33 s. The situation of AD is serious, but the precise physiological changes triggering the development of AD largely remain unknown. In this work, our method is used for a real AD dataset developed by National Alzheimer's Coordinating Center of US. The maximum of classification accuracies on the selected subsets arrives at 81.3 %. For the samples with AD, the maximum of classification accuracies is 82.75 %. These results verify our method is suitable tool to mine the characteristics of AD.

The rest of paper is organized as follows. Section 2 gives the background of our study. The basic knowledge about rough sets and mutual information is shown. The proposed feature selection scheme is detailed in Section 3. The experimental results are presented in Section 4. The application of AD dataset is in Section 5. Finally, the conclusion is drawn in Section 6.

2 Background

In this section, the related theories with our work are introduced. First, the concepts of mutual information is shown, which can be used to evaluate the similarities among features. Then, basic idea of rough sets and the extended method employing samples of the Boundary Region are given.

2.1 Mutual information

In according with Shannon's information theory (Cover and Thomas 1991), the uncertainty of a variable C can be measured by entropy $H(C)$. For two variables C and Y, the MI $I(C;Y)$ expresses the degree of reduced uncertainty about C after observing Y. It is calculated as

$$I\left(C;\,Y\right)=H\left(C\right)-H\left(\left.C\right|\,Y\right), \tag{1}$$

where $H\left(\left.C\right|\,Y\right)$ is conditional entropy measuring the uncertainty about C after observing Y.

When C is a discrete variable, with the entropy defined by Shannon, the entropy of C is expressed as

$$H\left(C\right)=-\sum_{c\in C}p\left(c\right)\log p\left(c\right), \tag{2}$$

where $p(\cdot)$ is the probability mass function of C. When $Y$is also a discrete variable, the conditional entropy $H(C|Y)$ is expressed as

$$H\left(\left.C\right|\,Y\right)=-\sum_{y\in Y}p\left(y\right)\left(\sum_{c\in C}p\left(\left.c\right|\,y\right)\log p\left(\left.c\right|\,y\right)\right), \tag{3}$$

where $p(c|y)$ represents the conditional probability mass of C and Y. Their MI is

$$I\left(C;\,Y\right)=\sum_{c\in C}\sum_{y\in Y}p\left(c,\,y\right)\log\frac{p\left(c,\,y\right)}{p\left(c\right)p\left(y\right)}, \tag{4}$$

where $p(c,\,y)$ is the joint probability mass function. When C and Y are continuous variables, the MI between C and Y is (Chow and Huang 2005)

$$I\left(C;\,Y\right)=\iint p\left(c,\,y\right)\log\frac{p\left(c,\,y\right)}{p\left(c\right)p\left(y\right)}dydc. \tag{5}$$

Based on the basic MI expression, quadratic MI is got by inserting quadratic expressions of the related variables (Torkkola and Campbell 2000), which are useful to reduce the complexity of computational cost. If D_1 and D_2 are discrete variables, their quadratic MI (Chow and Huang 2005; Huang and Chow 2005; Torkkola and Campbell 2000) is

$$I\left(D_1;\,D_2\right)=\log\frac{\sum\limits_{d_1\in D_1}\sum\limits_{d_2\in D_2}p\left(d_1,d_2\right)^2\sum\limits_{d_1\in D_1}\sum\limits_{d_2\in D_2}p\left(d_1\right)^2p\left(d_2\right)^2}{\left(\sum\limits_{d_1\in D_1}\sum\limits_{d_2\in D_2}p\left(d_1,d_2\right)p\left(d_1\right)p\left(d_2\right)\right)^2}. \tag{6}$$

For two continuous variables E_1 and E_2, their quadratic MI is:

$$I\left(E_1;\,E_2\right)=\log\frac{\iint p\left(e_1,e_2\right)^2 de_1de_2\iint p\left(e_1\right)^2 p\left(e_2\right)^2 de_1de_2}{\left(\iint p\left(e_1,e_2\right)p\left(e_1\right)p\left(e_2\right)de_1de_2\right)^2}. \tag{7}$$

The probability mass function describes the relative likelihood for the variable to occur at a given point, which is estimated according to the observation. Histogram and Kernel method are also used to estimate the probability mass function. However, Histogram tends to produce large estimation errors and pushes the memory requirement exponentially, with the increasing size of data.

2.2 Rough sets based feature selection

The theory of rough sets aims to approximately describe the sets that are unknown, incompletely specified, or whose specification is over complex. The fundamental notions of rough sets are lower and upper approximations of sets (Pawlak 1982, 1991; Pawlak and Skowron 2007a, b, c). The lower approximation is a description of the domain objects with certainty belonging to the concept of interest, whereas the upper approximation is a description of the domain objects that possibly belong to the concept of interest.

Definition 2.1 Let $IS(U,A)$ be a complete information system, where U is a nonempty finite set of objects and A is a nonempty finite set of features so that $f : U \rightarrow V_f$ for every $f \in A$. V_f is the set of values that f takes. For any $P \subseteq A$, there exists an indiscernible relation $IND(P)$

$$IND\left(P\right)=\left\{(x,y)\in U^2\left|\forall f\in P,\ f\left(x\right)=f\left(y\right)\right.\right\}. \tag{8}$$

Dataset can be seen as an information system, where samples are the objects of U and features are the elements of A.

Definition 2.2 A partition of U generated by p is defined as

$$U\big/IND\left(P\right)=\otimes\left\{f\in P, U\big/IND\left(\{f\}\right)\right\}, \tag{9}$$

where

$$\begin{array}{l} U\big/IND\left(\{f\}\right)=\left\{\{x|\,f\left(x\right)=b\,,x\in U\}|\,b\in V_f\right\} \\ A\otimes B=\left\{X\cap Y\left|\forall X\in A,\forall Y\in B,\ X\cap Y\neq\emptyset\right.\right\}\end{array}. \tag{10}$$

If $(x,y)\in IND\left(P\right)$, x and y are indiscernible according to the feature subset P. The equivalence class of x on the P-indiscernible relation is denoted by $[x]_p$. If x and y are indiscernible according to the features subset P, $y \in [x]_p$.Construct the P-lower approximations and P-upper approximations of X as

$$\underline{P}X=\left\{x\left|[x]_P\subseteq X\right.\right\}, \tag{11}$$

$$\overline{P}X=\left\{x\left|[x]_P\cap X\neq\emptyset\right.\right\}, \tag{12}$$

where (11) is the P-lower approximations, and (12) is the P-upper approximations.

By the definition of the P-lower approximations and P-upper approximations, the objects in U can be partition into three regions which are the Positive Region, the Boundary Region, and the Negative Region.

Definition 2.3 Positive Region, Boundary Region and Negative Region are defined as

$$POS_P(L) = \bigcup_{X \in U/IND(L)} \underline{P}X, \tag{13}$$

$$BND_P(L) = \bigcup_{X \in U/IND(L)} \overline{P}X - \bigcup_{X \in U/IND(L)} \underline{P}X, \tag{14}$$

$$NEG_P(L) = U - \bigcup_{X \in U/IND(L)} \overline{P}X. \tag{15}$$

Although L can be the set of any features, it is fixed as the set of label feature to reduce the burden of description. In this work, we just consider the system with one label feature. That is to say that $L = \{l\}$.

Definition 2.4 Dependency of label feature set L on a feature set P is calculated as

$$\gamma_P(L) = \frac{|POS_P(L)|}{|U|}, \tag{16}$$

where $|\cdot|$ is the number of objects in the set.

The following lists the QUICK REDUCT (Słowiński 1992) using forward search. C is the set of all conditional features. That is to say that C includes all features except label. R is the output of the algorithm. However, the output of QUICK REDUCT is possible to be not a reduct. For a set S, that is a subset of C, it is a reduct, if $POS_S(L) = POS_C(L)$ and $\forall a \in S,\ POS_{S-\{a\}}(L) \neq POS_S(L)$ (Pawlak 1991). The algorithm provides an addition strategy in constructing a subset. $POS_R(L) = POS_C(L)$ is verified by the value of dependency. However, $a \in R$ making $POS_{R-\{a\}}(L) = POS_R(L)$ possibly exists. In this case, the output of QUICK REDUCT is not a reduct. Three strategies shown in Yao et al. (2008) can make sure that the output is a reduct. They are reduction construction by deletion, reduction construction by addition-deletion and reduction construction by addition. Deletion strategy eliminates the features in the set of all conditional features until it finds a subset R producing the same Positive Region with the set of all conditional features. And there is not such feature $a \in R$ making $POS_{R-\{a\}}(L) = POS_R(L)$. This strategy is not efficient when a reduct is short (Yao et al. 2008). Addition-deletion strategy firstly constructs a set R by inserting features regarding to the Positive Region. Then, a deletion process reduces such feature $a \in R$ and $POS_{R-\{a\}}(L) = POS_R(L)$.

Reduction construction by addition inserts features by considering instinct matrix to directly get a reduct.

```
Algorithm 2.1: QUICK REDUCT
    Input: $C$ , the set of all conditional features
    Input: $L$ , the set of decisional feature
    Output: $R$ , a selected feature subset
                $R \leftarrow \{\}$
                    Repeat
                        $\forall f \in (C - R)$
                        if $\gamma_{R\cup\{f\}}(L) \succ \gamma_R(L)$
                            $R \leftarrow R \cup \{f\}$
                        end
                    Until $\gamma_R(L) = \gamma_C(L)$
                Return $R$
```

2.3 Boundary region of rough sets

QUICK REDUCT only considers the samples in the Positive Region, so the information of the samples in other regions can not be contained. Distance Metric Quick Reduct (DMQR) (Parthaláin et al. 2007) and Distance Metric Tolerance Rough sets (DM-TRS) (Parthaláin and Shen 2009) drive distance metrics as part of the feature selection criteria for finding a better feature subset. The distance metrics employed by DMQR and DM-TRS qualify the Objects in the Boundary Region with regard to their proximity to the low approximations. Since evaluating the margin of high dimensional space needs large computational effort, the two definitions calculate the mean of all samples in the P-lower approximations to reduce the computational burden. And the mean defined in Parthaláin et al. (2007) and Parthaláin and Shen (2009) is

$$\underline{P}X_{\text{mean}} = \left\{ \frac{\sum_{x\in\underline{P}X} f(x)}{\left|\underline{P}X\right|} |\forall f \in P \right\}. \tag{17}$$

Another definition (Parthaláin et al. 2010) of the mean is given as

$$\underline{P}X_{\text{mean}} = \left\{ \frac{\sum_{x\in\underline{P}X} f(x)}{|POS_P(X)|} |\forall f \in P \right\}. \tag{18}$$

Formulas (17) and (18) give two definitions of the mean. The means are employed to calculate the distance metric in formula (19). These two formulas calculate the means of samples in the Positive Region according to their explanation. However,

X is not described clearly. Moreover, they are general means of all objects. So the distance metric can not evaluate the clustering of samples in the same class. That is to say that the distance metric is not sure to be benefit to classification accuracy. In addition, it is more complex problem when the Positive Region is sparse, since there are not enough samples providing information.

The distance metrics of DMQR and DM-TRS are defined as

$$\omega_P(L) = \left(\sum_{y \in BND_p(L)} \delta\left(\underline{P}X_{\text{mean}}, y\right) \right)^{-1}, \tag{19}$$

where y is a sample in the Boundary Region, and δ is a distance function which is a Euclidean distance function described in Parthaláin et al. (2007) and Parthaláin and Shen (2009).

The feature selection criteria of DMQR and DM-TRS (Parthaláin et al. 2007; Parthaláin and Shen 2009) include two parts: the dependency and the distance metric of the samples in the Boundary Region. The evaluation measure of the two algorithms is defined as

$$M_P(L) = \frac{\omega_P(L) + \gamma_P(L)}{2}. \tag{20}$$

Apart from the feature selection criterion and stopping rule, their mechanisms are the same as QUICK REDUCT. Since $\gamma_R(L) = \gamma_C(L)$ is an ideal condition that cannot always be obtained in practical situations, DMQR and DM-TRS are set to stop when no new feature is found.

3 Class-based boundary rough sets with mutual information for feature selection

Rough sets based feature selection methods focus on analyzing the Positive Region which is the certain part of the data information. This, however, means the data information which lies in the Boundary Region is overlooked. DMQR and DM-TRS explore the distance metric of the samples in the Boundary Region as part of the feature selection criteria. Compared with other rough sets based feature selection methods, the feature subset determined by DMQR or DM-TRS exhibits a stronger relationship between selected features and labels. In this section, we introduce a new algorithm which selects features according to the data information of the Positive Region and Boundary Region. In the new method, a class-based distance metric is defined to measure the clustering degree of the samples in the Boundary Region and the Positive Region. With the class-based distance metric, samples in the Boundary Region are as close as possible to the samples having the same label in the Positive Region, so that the possibility of the samples in the Boundary Region decreases, which is classified incorrectly. Rough sets can not evaluate the degree of relevance improvement with a new feature inserted into the selected subset, so it can not always reduce the redundancy of features efficiently. In this paper, we use MI to eliminate feature redundancy. MI, which is robust to noise, can estimate the similarity of variables. Thus, the solution of the proposed method appears to be less redundant and more relevant.

3.1 Differentiated rough sets

For rough sets based FS methods, the dependency is one of the most important concepts. The dependency is based on samples in the Positive Region, so it verifies the definitive relevance between features and labels. DMQR and DM-TRS evaluate the uncertain relevance by the distance metrics which attempts to qualify the samples in the Boundary Region, with regard to their proximities to the samples in the P-low approximations. That is to say, the larger the distance metric is, the higher the likelihood that samples in the Boundary Region belong to the set of interest is Parthaláin and Shen (2009). But the P-lower approximations are not direct objects in the process of FS. Thus, in this paper, we focus on the Positive Region directly.

In some situations, the information concerning label feature is only partial. That is to say that the values of labels for some samples are missed. For rough sets, label feature is sensitive data, which plays important role in calculating relevance. When we categorize samples, missed labels must be considered.

We can categorize the samples of a dataset into six groups:

Type-1: For a sample, its equivalence class of P-indiscernible relation only includes itself. And its value of label is observed.

Type-2: For a sample, its equivalence class of P-indiscernible relation includes more than 1 object. Its value of label is observed. And all the samples in its equivalence class have the same value of label feature.

Type-3: For a sample, its equivalence class of P-indiscernible relation includes more than 1 object. The values of labels for the samples in its equivalence class are all observed. And there exists at least 2 samples in its equivalence class with different values of label feature.

Type-4: For a sample, its equivalence class of P-indiscernible relation only includes itself. And its value of label is not observed.

Type-5: The equivalence class of a sample includes more than 1 sample. There exits samples with observed values of labels and samples with missed labels in its equivalence class.

Type-6: The equivalence class of a sample includes more than 1 sample. All samples missed the values of label feature.

If there are some samples with missed labels, the dataset is not complete. In this work, we focus on complete dataset. That is to say that the values of all samples concerning label feature are observed.

Theorem 3.1 *The samples of Type-*1 *or Type-*2 *are in the Positive Region; the samples of Type-*3 *are in the Boundary Region.*

Proof of Theorem 3.1 When a sample "x" is of Type-1, the equivalence class $[x]_p$ of P-indiscernible relation only contains the sample itself. According to the definition of Type-1, the value of x concerning label is observed. So $\exists Z \in U\big/IND\,(L)$makes x $\in$ Z, where U/IND(L) is a partition of U generated by label feature. As a result, $[x]_p = \{x\} \subseteq Z$. According to Definition 2.3, the sample must be in the Positive Region.

When x is of Type-2, $[x]_p$ contains more than 1 samples. For label feature, the values of the samples in $[x]_p$are the same. So $\exists Z \in U\big/IND\,(L)$ and $[x]_p \subseteq Z$. Thus, x must be in the Positive Region.

When x is of Type-3, there exists at least 2 samples with different values of label feature. For $\forall Z \in U/IND(L)$, if $\exists y \in [x]_P$ makes y $\in Z$, $\exists y' \in [x]_P$ and $y' \notin Z$. So $x \notin \bigcup_{X \in U/IND(L)} \underline{P}X$. Since the values of labels for all samples in $[x]_p$are observed, $\exists Z' \in U/IND(L)$ makes $[x]_p \cap Z' \neq \emptyset$. As a result, x is in the Boundary Region. □

In rough sets, missed value is generally considered as any possible value in the same feature. However, this method is dangerous for label. First, the reason of missed value is that the value is not possible to be obtained or is definitively impossible to be obtained (Stefanowski and Tsoukias 2001). Second, this method decreases the possibility that samples in the Positive Region are classified into correct class. The samples of Type 4–6 can not be directly partitioned into regions before redefining $U/IND(L)$. Since incomplete system is not our objective in this work, its details are not given, and more contents of incomplete system can found in Stefanowski and Tsoukias (2001), Grzymala-Busse and Rzasa (2006) and Słowiński and Stefanowski (1989).

Table 1 shows an example elaborating the three types of samples and Theorem 3.1. The example contains 6 samples among which 3 samples are of class "F", 3 samples are of class "S". Sample 1 is of Type-1; sample 2 and 3 are of Type-2; sample 4–6 are of Type-3. So sample 1–3 are in the Positive Region; others are in the Boundary Region.

It is easy to verify each sample in complete dataset must be in only one group from Type-1 to Type-3, so Theorem 3.2 is found according to Theorem 3.1.

Theorem 3.2 *For complete information system, the samples in the Positive Region are of Type-1 or Type-2; the samples in the Boundary Region are of Type-3.*

In the Positive Region, we define that the samples, which are in the same equivalence class of P- indiscernible relation, consist of a cluster-element. The cluster-element of a Type-1 sample just contains the sample itself. Thus, the label of this cluster-element is the same with the sample. For a sample of Type-2, the samples in its cluster-element have the same label, so the cluster-element has the same label with the samples. By the analysis of cluster-elements, the Positive Region has the characteristic of labels. According to the concept of rough sets, the samples in the Positive Region are classified correctly. In theory, an optimal selected feature subset should be able to deliver high classification accuracy. In our work, we require that the samples in the Boundary Region should be with regard to their proximity to the samples in the Positive Region. Since the Positive Region has the characteristic of labels, we conduct partition of the samples in the Positive Region and Boundary Region according to their respective class labels.

Table 1 An example elaborating Theorem 3.1

Sample	Feature-1	Feature-2	Label	Sample	Feature-1	Feature-2	Label
1	2	3	S	4	2	4	F
2	4	5	F	5	2	4	S
3	4	5	F	6	2	4	S

Definition 3.1 Samples in the Positive region are divided into subsets where samples have the same class label. It can be denoted as

$$POSL_P = \left\{ POS_P(L) \cap Z \,\middle|\, Z \in U/IND(L) \right\}. \tag{21}$$

By Definition 3.1, the Positive Region is divided into several subsets according to label feature.

Definition 3.2 The mean set of elements in $POSL_P$ is defined as

$$POSL_{MP} = \{(Mean(PL)_P,\ L(PL)) |\, PL \in POSL_P\}, \tag{22}$$

where

$$Mean(PL)_P = \left\{ \frac{\sum\limits_{x \in PL} f(x)}{|PL|} \,|\forall f \in P \right\}. \tag{23}$$

$L(PL) = \{l(x)\,|\forall x \in PL\}$ and l is the label feature, since we just consider the system with unique decisional feature; $|PL|$ denotes the number of samples in the current PL. Since all the samples in one PL have the same label because of Definition 3.1, the $L(PL)$ of each object in $POSL_{MP}$ has only one element. Each object in $POSL_{MP}$ has two elements—the first one is the mean of the samples in the same PL, and the other is the label set.

Definition 3.3 Samples in the Boundary Region are divided into subsets where samples have the same class label. It can be denoted as

$$BNDL_P = \left\{ BND_P(L) \cap Z \,\middle|\, Z \in U/IND(L) \right\}. \tag{24}$$

Definition 3.4 Distance metric set is denoted as

$$DMS_P = \left\{ \sum_{x \in BL} \delta_P(x, y) \,\middle|\, BL \in BNDL_P, \left(y, L'\right) \in POSL_{MP}, L(BL) = L' \right\}, \tag{25}$$

where δ is a distance function, and $L(BL) = \{l(x)\,|\forall x \in BL\}$. $l(BL) = L'$ means that the elements of these two sets are completely the same. Since all the samples in one BL have the same label, the element of $L(BL)$ is unique.

Definition 3.5 Class-based distance metric on a feature set (CDM_P) is defined as

$$CDM_P = \sum_{\delta_P \in DMS_P} \delta_P^{-1}. \tag{26}$$

CDM_P is based on $POSL_P$and $BNDL_P$which are partitioned according to class labels. Thus, CDM_P exhibits the characteristic of labels. For samples with the same label, samples in the Boundary Region are getting closer to the mean of samples in the Positive Region by increasing CDM_P.That is to say that the uncertain part of dataset is getting closer to samples which are sure to be classified correctly. As a result, CDM_P evaluates the uncertain relevance of features and labels.

The combination of class-based distance metric and dependency is denoted as $CDM - D_P = CDM_P + \gamma_P$.

3.2 Forward feature selection process with CDM-D and MI

Rough sets based feature selection relies on an estimated relevance relationship between features and labels to find a feature subset. But the increasing degree of relevance by adding a new feature into subset is not considered. The *FSC* proposed in Chow and Huang (2005) can estimate the similarity between the feature subset *S* and a single feature f_m. The *FSC* is defined as

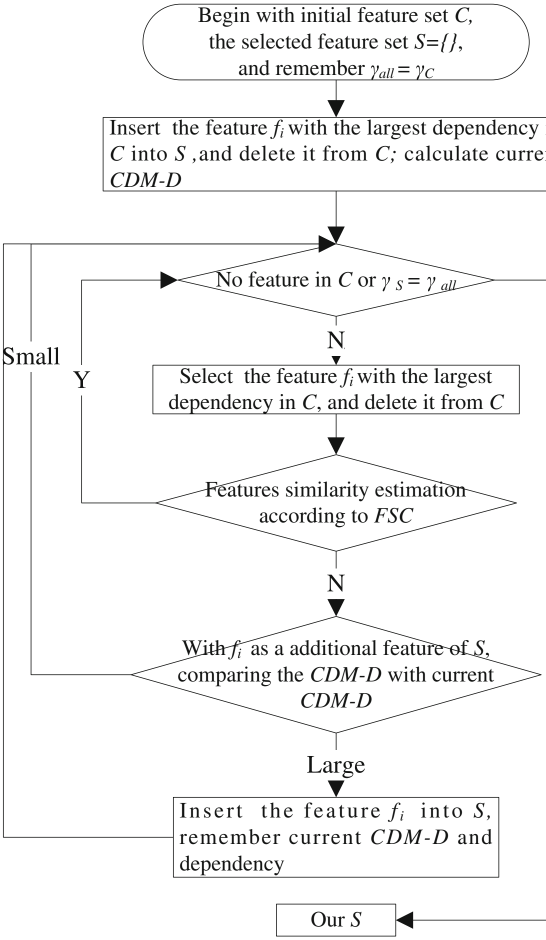
$$FSC\,(f_m) = \arg\max_{f_i \in S} \left(\frac{I\,(f_m;\, f_i)}{H\,(f_i)} \right). \tag{27}$$

When $FSC(f_m)$ is large enough, i.e, $FSC\,(f_m) \geq \theta$, the feature f_m can be considered as a redundant feature for *S*, and should not be added into *S*. Throughout this paper, θ is set to 0.95. In order to verify the effectiveness of similarity estimation, the comparison of $CDM - D$ on different feature subsets in the LED dataset, which is highly redundant, is presented in Section 4. Without the *FSC*, the redundant features are inserted into the selected subset, so the increasing of $CDM - D$ is slow. In contrast, the redundant features can be reduced by *FSC*, enabling the $CDM - D$ reaches its own maximum in a much speedy way.

The proposed RSMI-FS is a forward process. It begins with an empty feature set, and additional features are included in the way of one by one. The process of FS is guided by the dependency of each feature. That is to say the initial feature sequence is determined by its own dependency. Based on $CDM - D$ and *FSC*, the RSMI-FS algorithm is realized as follow, and its flow diagram is illustrated in Fig. 1.

```
Algorithm 3.1: Feature selection method based on rough sets and mutual information (RSMI-FS)

Step 1) Set $C$ , $S$  the conditional feature set and an empty set, respectively. Remember $\gamma_{all} = \gamma_C$ .

Step 2) In $C$ , find out the feature  $f_i$  having the maximal dependency. Put  $f_i$ into $S$ , and delete it
          form $C$ . Remember the current $CDM - D = CDM - D_{\{f_i\}}$ and current dependency $\gamma_{\{f_i\}}$ .

Step 3) If $\gamma_S = \gamma_{all}$ , or there is no features remaining in $C$ , go to step 6.

Step 4) Select the feature $f_i$ having the maximal dependency, and delete it form $C$ .

Step 5) If $FSC\{f_i\} < \theta$  and  $CDM - D_{\{f_i\} \cup S} \succ CDM - D_S$ , put $f_i$  into $S$ , and remember the current
        $CDM - D = CDM - D_S$  and dependency $\gamma_S$ . Then, go to step 3.

Step 6) Output $S$ .
```

Fig. 1 Flow diagram of RSMI-FS


Begin with initial feature set C, the selected feature set $S=\{\}$, and remember $\gamma_{all} = \gamma_C$
Insert the feature f_i with the largest dependency in C into S ,and delete it from C; calculate current $CDM\text{-}D$
No feature in C or $\gamma_S = \gamma_{all}$
N
Select the feature f_i with the largest dependency in C, and delete it from C
Features similarity estimation according to FSC
N
With f_i as a additional feature of S, comparing the $CDM\text{-}D$ with current $CDM\text{-}D$
Large
Insert the feature f_i into S, remember current $CDM\text{-}D$ and dependency
Our S
Small
Y
Y

4 Results and discussion

In this study, 2 synthetic datasets and 6 real datasets were used. The 2 synthetic datasets verify the correctness of the proposed RSMI-FS. The first synthetic dataset highlights the advantage of computational time of finding an optimal subset by RSMI-FS. The highly redundant LED domain dataset is used to verify the capability of dealing with the redundancy and irrelevance in feature set. For 6 real datasets, five indexes, namely Naïve Bayes, NNge, DTNB, OneR, and JRIP, are adopted to evaluate classification accuracies of the selected subsets. A comparative analysis among RSMI-FS, QUICK REDUCT, and DMQR, and DM-TRS is based on subset size and classification accuracy. Additionally, for RSMI-FS, the difference between Mahalanobis Distance and Euclidean Distance is discussed. Then, three MI based methods (FS-RAW-MI (Bonnlander 1996), PWMI (Kwak and Choi 2002b), and OFS-MI (Chow and Huang 2005)), two fuzzy rough methods (Attribute Selection with Fuzzy Decision Reducts (Cornelis et al. 2010a), and Vaguely Quantified Rough Sets based Method (VQRS) (Cornelis and Jensen 2008)), and typical Filter methods

(Relief (Kononenko 1994; Robnik-Sikonja and Kononenko 1997), and Filter method based on correlation-based feature subset selection (Filter-CFS) (Hall 1998)) are also compared with RSMI-FS.

4.1 Feature subset evaluation index

In this section, five classifiers are employed to evaluate the classification accuracies of the selected subsets. They are Naïve Bayes, NNge, DTNB, OneR, and JRIP which are briefly outlined below.

Naïve Bayes (John and Langley 1995) is based on Bayesian Classifier. For Bayesian network, the model of probability mass function is an important problem, when the variables are continuous. Most works solve the problems by discretizing, or assuming that the data are generated by Gaussian distribution. But this assumption is not suitable for certain domains, such as clear semantics. Naïve Bayes employs a kernal estimation to approximate more complex distribution.

NNge (Martin 1995) is an instance-based learning classifier that classifies new examples by comparing them to those already known. A main problem of instance-based learners is that running time increases as more examples are used for learning. NNge proposes a non-nested generalized exemplar to solve this problem, and represents more useful rules. With the generalized exemplar, NNge reduces the role of distance function to determine the class, and decreases the classification errors caused by the inaccuracies of distance function.

The algorithm for learning the combined model (DTNB) (Hall and Frank 2008) investigates a semi-Naïve Bayesian ranking method that combines Naïve Bayesian with induction of decision table. DTNB splits the set of features into two groups. The class probability of one group is assigned according to Naïve Bayesian, the other group is based on decision table, and the resulting probability estimation is combined.

The OneR algorithm (Holte 1993) is based on the theory that highly accurate results on most datasets can be obtained by simple rules. The OneR uses a system, called 1R, whose input is a set of training examples and whose output is a 1-rule. OneR constructs a relatively small set of candidate rules, and selects one of these rules. It has been verified that, on many datasets, the performance of OneR is highly competitive with some complex numerous implications in the convex of machine learning research and applications.

JRIP (Cohen 1995) learns propositional rules by growing rules and pruning them to produce error reduction. Before a termination condition is satisfied, antecedents are inserted greedily during the growing phase. Then, the antecedents are pruned according to a pruning metric. When the rule set is generated, an optimization is performed to evaluate the rules.

4.2 Synthetic data set of varying size

Four dimension synthetic datasets $F = \{f_1, f_2, f_3, f_4\}$ are generated in this experiment. These datasets consist of 100, 500, 1,000, 2,000, 3,000, 4,000 samples, respectively. For each dataset, the samples belong to two classes $\{1, 2\}$, and each class has one half of the samples. For the input variables f_1 and f_2, the data is generated from the following two Binomial Distributions:

Class 1: (half samples of the dataset) $\{f_1, f_2\} - \text{binornd}\left([10, 10], [0.5, 0.5]\right)$
Class 2: (half samples of the dataset) $\{f_1, f_2\} - \text{binornd}\left([20, 20], [0.5, 0.5]\right)$.

The input variable f_3 is equal to the sum of f_1 and f_2, and f_4 is equal to $2 \times f_2$. Obviously, the input variables f_1 and f_2 are considered more important than f_3 and f_4 in this experiment. So a good feature selection method should select the subset $\{f_1, f_2\}$.

Table 2 shows the selection results of the compared methods. In this section, RSMI-FS uses Euclidean distance to calculate $CDM - D$. It indicates that RSMI-FS, ORS, and Relief are able to identify the relevant features correctly. ORS (Bazan et al. 2000), which can also find an optimal method, is introduced to compare the computational time with RSMI-FS. Relief can find the correct subset, but it needs the priori knowledge that 2 features are suitable. QUICK REDUCT finds the correct subset when the number of samples is 1,000. DMQR finds the correct subsets when the numbers of samples are 500 and 2,000. Also, it is noticed that DM-TRS with tolerance value 0.9, FS-RAW-MI, PWMI, and OFS-MI are not able to obtain the correct results in each dataset. For Fuzzy Decision Reducts method, $\alpha = 0.95$ is set in this work, which is suitable choice verified in Cornelis et al. (2010a). For VQRS based method, 0 and 0.8 are employed to define the VQRS Positive Region (Cornelis and Jensen 2008). The values of current $CDM - D - D$ on different numbers of samples are shown in Fig. 2. When the selection process stops according to its stopping criterion, additional features are not selected into the subset by RSMI-FS. Although the remaining features are not deleted from the initial feature set, the values of current dependency and $CDM - D - D$ do not change after the process stops. We find that, with the increasing of sample number, the maximum of related $CDM - D - D$ in Fig. 2 decreases. This is because that the number of samples, which have the same feature value and different labels, is likely to be larger when the dataset is larger.

Table 2 Results of feature selection on the synthetic dataset

Methods	Sample number					
	100	500	1,000	2,000	3,000	4,000
Quick reduct	$\{f_2, f_3\}$	$\{f_1, f_4\}$	$\{f_1, f_2\}$	$\{f_3, f_3\}$	$\{f_2, f_4\}$	$\{f_1, f_4\}$
DMQR	$\{f_4, f_2\}$	$\{f_2, f_1\}$	$\{f_3, f_4\}$	$\{f_2, f_1\}$	$\{f_1\}$	$\{f_4, f_1\}$
DM-TRS	$\{f_1, f_3\}$	$\{f_2, f_3\}$	$\{f_4, f_1, f_3\}$	$\{f_3\}$	$\{f_4, f_3, f_1\}$	$\{f_2, f_3\}$
ORS	$\{f_1, f_2\}$					
FS-RAW-MI	$\{f_1, f_2, f_4\}$					
PWMI	$\{f_1, f_2, f_4\}$					
OFS-MI	$\{f_1, f_2, f_3, f_4\}$	$\{f_1, f_2, f_4\}$				
Fuzzy decision reducts	$\{f_1, f_3, f_4\}$	$\{f_1, f_2, f_3, f_4\}$			–	–
VQRS	$\{f_1, f_3, f_4\}$	$\{f_1, f_4\}$	$\{f_1, f_2, f_3, f_4\}$		–	–
Relief	$\{f_1, f_2\}$					
Filter-CFS	$\{f_1, f_4\}$	$\{f_1, f_4\}$	$\{f_1, f_{24}\}$	$\{f_1, f_4\}$	$\{f_1, f_4\}$	$\{f_1, f_4\}$
RSMI-FS	$\{f_1, f_2\}$					

Fig. 2 CDM-D-D of the datasets with increasing samples. |samples| expresses the number of samples in the dataset

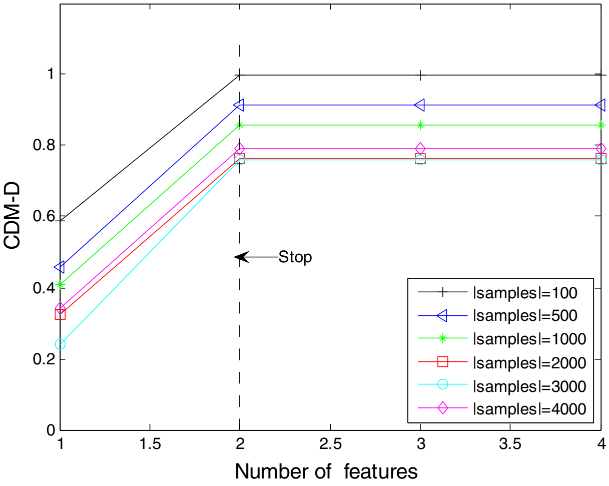
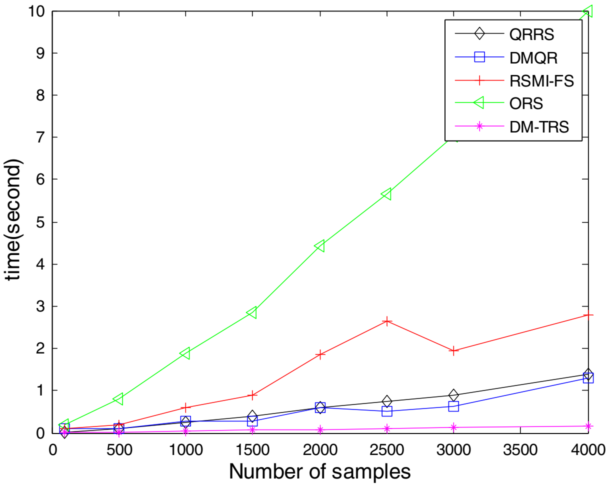


Figure 3, which illustrates the average computational time on different numbers of samples for 5 trials, shows that the computational time cost of RSMI-FS is not significantly larger than QUICK REDUCT, DMQR, and DM-TRS. ORS can find the correct feature subset for datasets with different number of samples, but its time cost is substantially larger compared with other four rough sets methods. Because each new feature found by ORS must most dramatically increase the dependency of its combination with the features having been in selected set. That is to say that all features not in selected set must be considered in each selecting step. As a result, the feature number plays significant role in computation cost of time, which is illustrated in Fig. 4. In our study, with the number of samples being set at 1,000, the features ranged from 4 to 20 are used for conducting a running time comparative analysis. The additional features are generated by the linear combination of f_1 and f_2 which

Fig. 3 Comparison of computational time cost with increasing number of samples



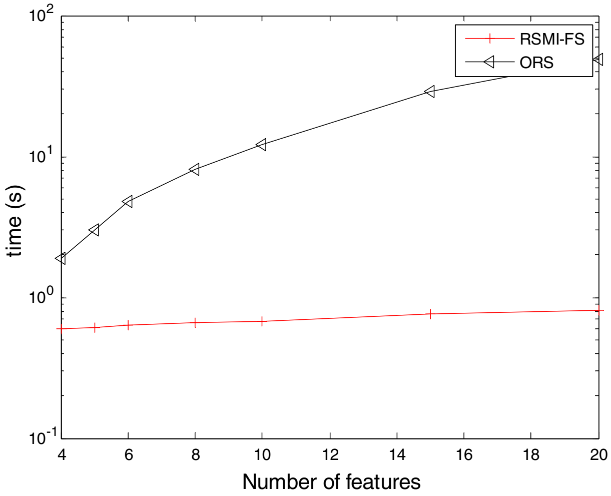


Fig. 4 Computational time cost with increasing features

is similar with f_3 or f_4. In Fig. 4, with the increasing of feature numbers, the running time of ORS increases dramatically.

4.3 Highly redundant LED display domain dataset

The dataset of LED display Domain has 24 features in which the first 7 features determine the concept of LED display, whilst the rest 17 features are redundant (Aha 1992). In this work, 1,000 samples are generated. With the priori knowledge, a subset, which consists of the first 7 features, is a good result. Table 3 shows the results of the first 7 features in the feature subsets found by the FS methods. In Table 3, f_i represents the feature in the selected feature subset, and (f_i) represents the feature which is not included by feature selection. There are other features in the results, so QUICK REDUCT, DMQR, and DM-TRS cannot avoid the redundancy in the selected feature sets. MI-based Method, Fuzzy Decision Reducts and Filter-CFS can deal with redundancy of the feature set, but the relevance features are also ruled out from the selected subsets. On the contrary, RSMI-FS, Relief, and VQRS find the correct feature subset which only contains the first 7 features.

Table 3 Results of feature selection on LED dataset

Feature selection method	Result of feature selection
Quick reduct	$\{(f_1), f_2, f_3, f_4, f_5, (f_6), (f_7)\}$
DMQR	$\{(f_1), (f_2), (f_3), (f_4), (f_5), f_6, (f_7)\}$
DM-TRS	$\{(f_1), (f_2), (f_3), (f_4), (f_5), (f_6), f_7\}$
FS-RAW-MI	$\{f_5, f_2, f_3, f_4, f_1, f_6, (f_7)\}$
PWMI	$\{f_5, f_2, f_3, f_4, f_1, f_6, (f_7)\}$
OFS-MI	$\{f_5, f_2, f_3, f_4, f_1, f_6, (f_7)\}$
Fuzzy decision reducts	$\{f_1, f_2, f_3, f_4, f_5, (f_6), (f_7)\}$
VQRS	$\{f_1, f_2, f_3, f_4, f_5, f_6, f_7\}$
Relief	$\{f_2, f_5, f_7, f_4, f_3, f_1, f_6\}$
Filter-CFS	$\{f_1, f_2, f_3, f_4, f_5, (f_6), (f_7)\}$
RSMI-FS	$\{f_6, f_3, f_5, f_4, f_1, f_7, f_2\}$

In order to show the necessary of MI for RSMI-FS, the increasing value of current $CDM - D - D$ with the number of features is shown in Fig. 5. When the selection progress of RSMI-FS stops, all the samples are in the Positive Region. Hence, the related maximum of $CDM - D - D$ in Fig. 5 is 1. When feature similarities are not evaluated by MI, 15 features are selected in order to partition all samples into the Positive Region. In contrast, only 7 features are in the subset found by RSMI-FS, which is shown in Table 3. This comparative analysis verifies that MI is necessary for RSMI-FS.

4.4 Comparison of RSMI-FS with rough sets based methods

This section shows the results of experimental studies using 5 UCI datasets and one YALE datasets (Georghiades et al. 2001). The data in Table 4 show the classification accuracies using the 5 classifiers, which are expressed as percentage. The setup of the experiments in this work is the same with Parthaláin and Shen (2009); Parthaláin et al. (2010), where classification using 10-fold cross validation is initially performed on the unreduced dataset, following by the reduced datasets. In this section, the results of RSMI-FS with Euclidean distance and Mahalanobis distance are shown to evaluate the effect of different distance functions. The results of DM-TRS setting tolerance values 0.8 or 0.9 are both considered.

In Table 4, the classification accuracies of RSMI-FS are larger or comparable with the related unreduced set. This verifies that FS is useful to find an indicative subset and remove measurement noise. For Small Soybean, Heart and Breast Tumor Diagnosis, the performance of RSMI-FS is almost better than other methods. Especially, for Small Soybean and Heart, the subsets of RSMI-FS do the best for every classifier. Although the performances of RSMI-FS and QUICK REDUCT are the same in Small Soybean, the features in the subset selected by QUICK REDUCT is 2.5 times as RSMI-FS, which is shown in Table 5. Comparing with other methods, their improvement of classification is more than 21.2766 %. For YALE,

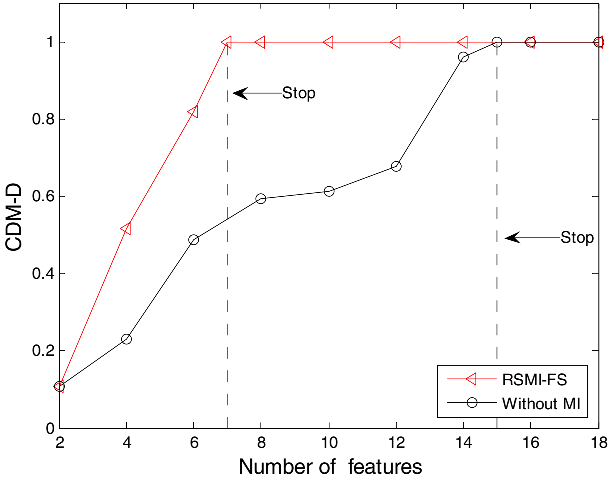


Fig. 5 Comparing of $CDM - D - D$ between RSMI-FS and the method without MI

Table 4 Classification accuracies of FS methods based on rough sets

Classifier (%)	FS method						
	Unreduced set	RSMI-FS (Euclidean distance)	RSMI-FS (Mahalanobis distance)	Quick reduct	DMQR	DM-TRS (tolerance value 0.9)	DM-TRS (tolerance value 0.8)
Data set	Small soybean (47 samples and 35 features)						
Naïve Bayes	97.8723	100	100	100	72.3404	40.4255	36.1702
DTNB	100	100	100	100	74.4681	36.1702	36.1702
OneR	82.9787	82.9787	82.9787	82.9787	57.4468	36.1707	36.1702
NNge	97.8723	100	100	100	78.7234	25.5319	27.6596
JRIP	97.8723	100	100	100	78.7234	36.1702	36.1702
	SPECT (267 samples and 21 features)						
Naïve Bayes	75.2809	79.4007	79.4007	79.4007	79.4007	79.4007	79.4007
DTNB	76.03	79.4007	79.4007	79.4007	79.4007	79.4007	79.4007
OneR	79.4007	79.4007	79.4007	79.4007	79.4007	79.4007	79.4007
NNge	76.4045	71.161	71.161	70.0375	64.794	65.9176	63.6704
JRIP	79.0262	79.4007	79.4007	79.4007	79.4007	79.4007	79.4007
	SPECTF (267 samples and 44 features)						
Naïve Bayes	68.5393	67.0412	70.0375	69.6696	67.4157	70.412	70.0375
DTNB	76.03	81.2734	81.2734	75.6554	77.1536	70.7865	79.0262
OneR	75.6554	80.5243	80.5243	76.4045	77.9026	77.9026	77.9026
NNge	79.4007	76.03	74.1573	74.1573	73.7828	78.6517	79.0262
JRIP	75.6554	79.0262	77.1536	80.5243	78.2772	73.7828	80.1498
	Heart (294 samples and 13 features)						
Naïve Bayes	82.9932	77.8912	77.8912	77.665	76.5306	73.8095	76.1905
DTNB	84.3537	82.3129	82.3129	79.932	78.5714	76.5306	78.5714
OneR	78.5714	81.9728	81.9728	81.2925	78.5714	78.5714	78.5714
NNge	77.551	77.8912	77.8912	76.1905	77.2109	72.7891	76.8707
JRIP	79.5918	81.2925	81.2925	79.932	78.5714	77.2109	78.5714
	Breast tumor (699 samples and 9 features)						
Naïve Bayes	95.9943	96.5616	96.7096	96.5616	95.9885	95.442	87.2675
DTNB	96.9957	95.702	96.4235	94.8424	95.9885	96.4235	90.701
OneR	92.7039	92.2636	92.7039	91.9771	91.1175	89.9857	90.701
NNge	96.2804	95.702	95.422	94.8424	94.5559	94.8498	87.6967
JRIP	95.1359	95.5651	94.9928	94.8498	94.4206	94.7067	90.701
	YALE (2,415 samples and 1,024 features)						
Naïve Bayes	–	53.604	53.604	52.9412	52.5269	54.7639	53.0655
DTNB	–	60.3977	60.3977	57.5808	57.208	67.8128	56.4209
OneR	–	61.599	61.599	51.8641	52.9826	57.008	54.6396
NNge	–	–	–	–	–	–	–
JRIP	–	65.7001	65.7001	60.0249	58.9478	81.2345	59.942

the best performance is found by DM-TRS (tolerance value 0.9), but 86 features are selected into the subset, which is 13 times larger than the results of RSMI-FS. The computational time cost of DM-TRS is 3827.3 s, but the cost of RSMI-FS with Euclidean distance and Mahalanobis distance is just 396.4 and 391.7 s, respectively.

Table 5 Numbers of features selected by rough sets based methods

Dataset	FS method					
	RSMI-FS (Euclidean distance)	RSMI-FS (Mahalanobis distance)	Quick reduct	DMQR	DM-TRS (tolerance value 0.9)	DM-TRS (tolerance value 0.8)
Small soybean	2	2	5	4	1	2
SPECT	2	2	3	2	1	1
SPECTF	3	2	4	7	19	3
Heart	6	6	10	3	6	5
Breast tumor	4	6	7	5	5	1
YALE	6	6	7	3	86	6

For FS methods based on rough sets, their dependencies evaluate the definitive relevance of features with labels. Figure 6 shows the values of the current dependency with increasing features. For rough sets based method, when the selection process stops according to the stopping criterion, additional features are not selected. Although remaining features are not deleted from the initial feature set, they are not useful to increase the dependency. In Fig. 6, RSMI-FS (ED) and RSMI-FS (MD) express that RSMI-FS uses Euclidean distance and Mahalanobis distance to calculate distance metric, respectively. Figure 6 shows that the maximal dependencies of RSMI-FS are always larger than DMQR and DM-TRS. Comparing with QUICK REDUCT setting the dependency as its selection criterion, RSMI-FS can reach its maximum value in a faster rate. In SPECT, the dependency of QUICK REDUCT is larger than RSMI-FS, but its accuracies shown in Table 4 are not larger than RSMI-FS. According to those results, the features selected by RSMI-FS are more relevant and less redundant.

In this work, we use Euclidean distance and Mahalanobis distance, which are widespread in machine learning, to calculate $CDM - D - D$ of RSMI-FS. Their difference of classification accuracies shown in Table 4 ranges from 0 to 2.99 %. Figure 7 shows the values of current $CDM - D - D$ with the increasing features. It can be observed that the difference of $CDM - D - D$ is not distinctive. As a result, we can conclude that, in RSMI-FS, the difference between Euclidean distance and Mahalanobis distance is not significant.

4.5 Comparison of RSMI-FS with MI based methods

In this section, the results of RSMI-FS are compared with other three MI based FS methods which are FS-RAW-MI, PWMI, and OFS-MI. Table 6 shows their classification accuracies. For Small Soybean, Heart, and Breast Tumor, the accuracies of the subsets found by RSMI-FS with the 2 distance functions are almost larger than the other methods. For Heart, the accuracies of PWMI in Naïve Bayes and DTNB are larger than RSMI-FS. But the number of features found by PWMI is 13, which is shown in Table 7. That is to say that PWMI selects all features of Heart. So PWMI does not play role in this dataset. For YALE, the performance of MI based methods is better than RSMI-FS, but the difference is insignificant. RSMI-FS selects the most relevant features, so its number of selected features is rather smaller than

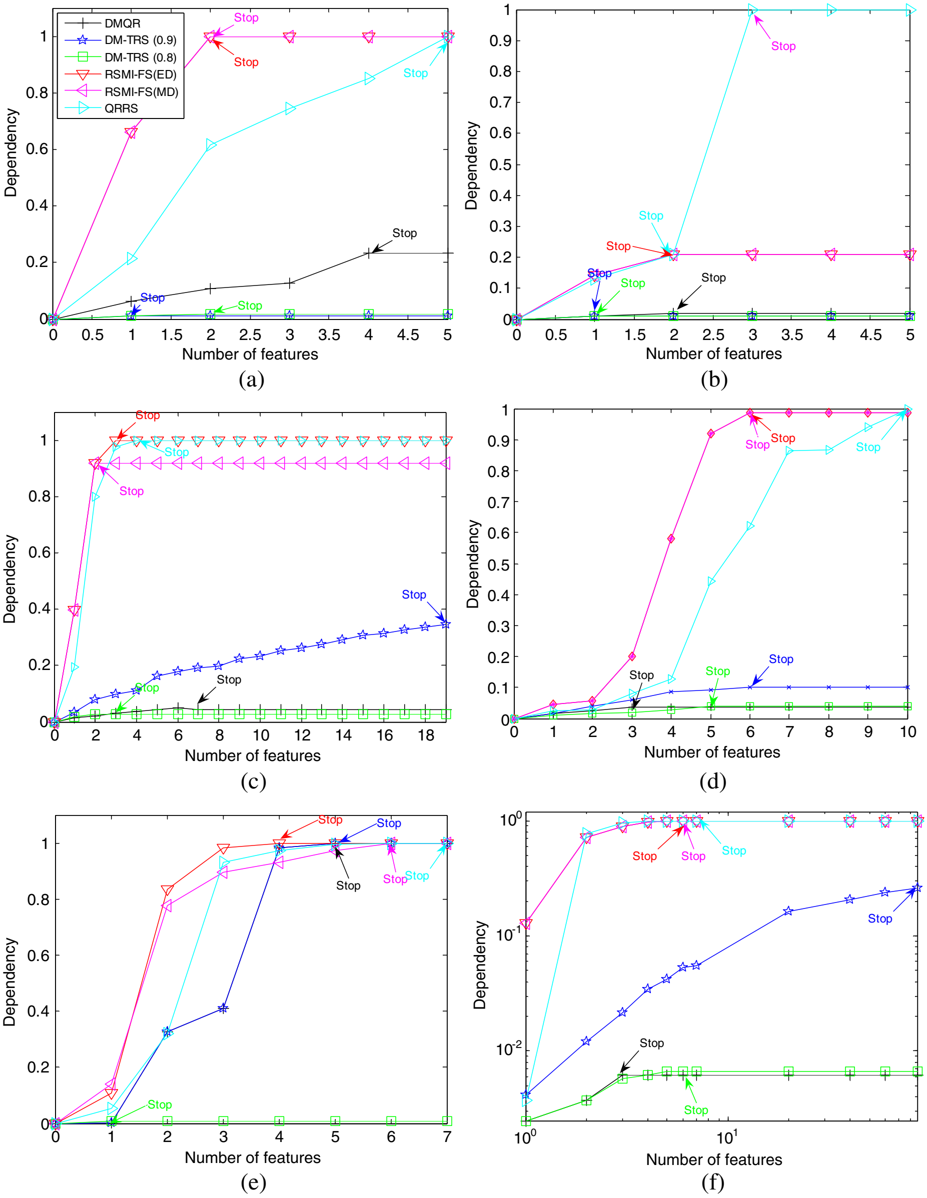
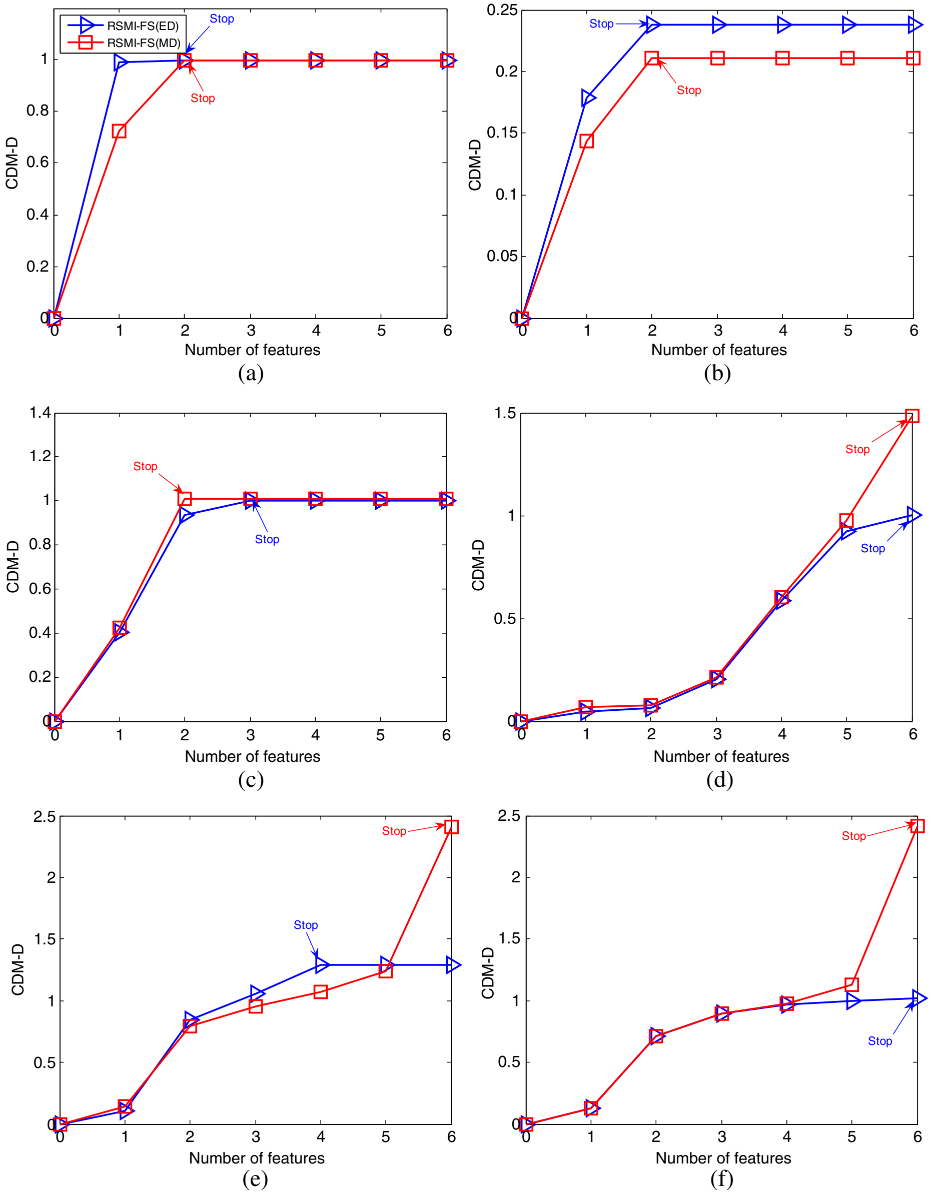


Fig. 6 Values of dependencies with increasing features, **a** Small Soybean **b** SPECT **c** SPECTF **d** Heart **e** Breast Tumor **f** YALE

MI methods. Moreover, the computational time cost of RSMI-FS is significantly less than the other three methods, and the difference of computational time cost ranges from 9,504 s to 24,706 s.



(a) (b) (c) (d) (e) (f)

Fig. 7 Values of $CDM-D-D$ with increasing features, **a** Small Soybean **b** SPECT **c** SPECTF **d** Heart **e** Breast Tumor **f** YALE

The numbers of features found by RSMI-FS are consistently less than the MI methods, which are shown in Table 7. And, in terms of classification accuracies detailed in Table 6, RSMI-FS are larger or comparable. The above findings in Sections 4.4 and 4.5 confirm RSMI-FS is able to find a more relevant and less redundant feature subset, compared with the rough sets or MI based FS methods.

Table 6 Classification accuracies of RSMI-FS and MI based methods

Classifier (%)	FS method				
	RSMI-FS (Euclidean distance)	RSMI-FS (Mahalanobis distance)	FS-RAW-MI	PWMI	OFS-MI
Data set	Small soybean				
Naïve Bayes	100	100	97.8723	95.7447	100
DTNB	100	100	100	95.7447	100
OneR	82.9787	82.9787	78.7234	55.3193	82.9787
NNge	100	100	97.8723	97.8723	97.8723
JRIP	100	100	100	95.7447	97.8723
	SPECT				
Naïve Bayes	79.4007	79.4007	76.4045	75.6554	77.5281
DTNB	79.4007	79.4007	77.1536	76.779	79.0262
OneR	79.4007	79.4007	79.4007	79.4007	79.4007
NNge	71.161	71.161	76.779	75.6554	79.4007
JRIP	79.4007	79.4007	80.5243	81.2743	82.397
	SPECTF				
Naïve Bayes	67.0412	70.0375	78.2772	71.5356	68.6593
DTNB	81.2734	81.2734	78.6517	77.1536	76.03
OneR	80.5243	80.5243	77.9026	79.4007	75.6554
NNge	76.03	74.1573	79.7753	73.4082	79.4007
JRIP	79.0262	77.1536	79.0262	78.2772	75.6554
	Heart				
Naïve Bayes	77.8912	77.8912	77.2727	82.9932	79.5918
DTNB	82.3129	82.3129	80.5924	84.3537	77.2109
OneR	81.9728	81.9728	81.2925	78.5714	81.2925
NNge	77.8912	77.8912	78.5714	77.551	76.5306
JRIP	81.2925	81.2925	76.5306	79.5918	78.5714
	Breast tumor				
Naïve Bayes	96.5616	96.7096	95.4155	95.1289	95.1289
DTNB	95.702	96.4235	95.1289	94.8424	94.8424
OneR	92.2636	92.7039	92.2636	91.9771	91.9771
NNge	95.702	95.422	95.702	93.9828	93.9828
JRIP	95.5651	94.9928	95.7082	95.9943	95.9943
	YALE				
Naïve Bayes	53.604	53.604	57.9122	56.7523	58.575
DTNB	60.3977	60.3977	66.6259	67.4399	–
OneR	61.599	61.599	59.6106	61.7647	61.3505
NNge	–	–	–	–	–
JRIP	65.7001	65.7001	78.2461	79.2461	79.8674

4.6 Comparison of RSMI-FS with other methods

In this section, RSMI-FS is compared with two fuzzy rough methods and two typical non rough sets methods. Fuzzy sets and rough sets are two natural computing methods to deal with data of inconsistency and uncertainty in a human-like fashion

Table 7 Numbers of features found by RSMI-FS and MI methods

Dataset	FS method				
	RSMI-FS (Euclidean distance)	RSMI-FS (Mahalanobis distance)	FS-RAW-MI	PWMI	OFS-MI
Small soybean	2	2	3	5	13
SPECT	2	2	18	20	17
SPECTF	3	2	3	3	44
Heart	6	6	10	13	3
Breast tumor	4	6	3	3	3
YALE	6	6	17	23	40

(Jensen and Cornelis 2011b). Their difference is the type of uncertainty and their approach to deal with it (Jensen and Cornelis 2011a). In this work, Attribute Selection with Fuzzy Decision Reducts (Cornelis et al. 2010a) and Vaguely Quantified Rough Sets based Method (VQRS) (Cornelis and Jensen 2008) are compared with RSMI-FS. By setting the degree of reducthood α, the subset found by Fuzzy Decision Reducts can provide a comparable accuracy with small size of features. In this work, α is set 0.95 which is suitable value introduced in Cornelis et al. (2010a). The other details of setup about this method are the same with Cornelis et al. (2010a). Fuzzy rough methods are abrupt in a sense that adding or omitting a single element may drastically alter the outcome of approximations. So any misclassified object prevents rough sets from making any conclusive statement about all objects related to it. VQRS based method defines a smoother region of tolerance towards classification errors to reduce this kind negative impact. The parameters of the region are set 0 and 0.8 which is employed in Cornelis and Jensen (2008). Relief (Kononenko 1994; Robnik-Sikonja and Kononenko 1997) gives each feature a relevance weighting that reflects its ability to discern class labels. It is typically used in conjunction with a feature ranking method to select features. In order to comparing the performance with our method, the number of its selected features is set as the same with RSMI-FS. When the numbers of features selected by RSMI-FS with two distance functions are different, the number for Relief is set as the larger one. For Filter method based correlation-based feature subset selection (Hall 1998) (Filter-CFS) runs a subset evaluator on the data passed through a resample filter. The subset evaluator is correlation-based feature subset selection which evaluates the worth of a feature subset by considering the individual predictive ability of each feature along with the degree of redundancy among the features. Filter-CFS employs linear forward search (Guetlein et al. 2009) which is an extension of best first.

Table 8 shows the accuracy performance. For Small Soybean, the accuracies of subsets found by RSMI-FS, VQRS, and Fuzzy Decision Reducts are the same in Naïve Bayes, DTNB, NNge, and JRIP. In OneR, the accuracy of RSMI-FS is 4.26 % more than VQRS and Fuzzy Decision Reducts. For the same dataset, comparing with Relief, the improvement of RSMI-FS is more than 23.4 %. For SPECTF, RSMI-FS obtain the largest classification accuracies in DTNB and OneR. For the results of JRIP about SPECTF, RSMI-FS with Euclidean distance is better than other methods. Moreover, features selected by RSMI-FS are not more than others, shown

Table 8 Classification accuracies of RSMI-FS and other methods

Classifier (%)	FS method					
	RSMI-FS (Euclidean distance)	RSMI-FS (Mahalanobis distance)	VQRS	Fuzzy decision reducts	Relief	Filter-CFS
Data set	Small soybean					
Naïve Bayes	100	100	100	100	76.5957	100
DTNB	100	100	100	100	76.5957	97.8723
OneR	82.9787	82.9787	78.7234	78.7234	57.4468	82.9787
NNge	100	100	100	100	70.2128	97.8723
JRIP	100	100	100	100	55.3191	97.8723
	SPECT					
Naïve Bayes	79.4007	79.4007	79.4007	75.2809	76.4045	77.9026
DTNB	79.4007	79.4007	79.4007	76.4045	79.4007	79.7753
OneR	79.4007	79.4007	79.4007	79.4007	79.4007	79.4007
NNge	71.161	71.161	67.0412	76.779	71.9107	77.9026
JRIP	79.4007	79.4007	79.4007	80.5234	79.4007	76.9231
	SPECTF					
Naïve Bayes	67.0412	70.0375	66.6667	70.7865	75.1685	70.7865
DTNB	81.2734	81.2734	79.4007	81.2734	79.4007	74.5318
OneR	80.5243	80.5243	78.2772	77.5281	76.03	74.5318
NNge	76.03	74.1573	77.9026	75.6554	73.0337	80.1498
JRIP	79.0262	77.1536	78.2772	74.1573	74.1573	77.5281
	Heart					
Naïve Bayes	77.8912	77.8912	74.4898	77.551	81.2925	82.3139
DTNB	82.3129	82.3129	78.5714	78.2313	80.2721	79.2517
OneR	81.9728	81.9728	81.9728	81.9728	81.2925	81.2925
NNge	77.8912	77.8912	74.8299	78.5714	77.8912	73.1293
JRIP	81.2925	81.2925	82.9932	82.9932	79.5918	77.2109
	Breast tumor					
Naïve Bayes	96.5616	96.7096	95.9943	96.1373	96.4235	95.9943
DTNB	95.702	96.4235	97.1388	95.8512	97.2818	96.9957
OneR	92.2636	92.7039	92.7039	92.7039	92.7039	92.7039
NNge	95.702	95.422	96.7096	96.2804	96.2804	96.2804
JRIP	95.5651	94.9928	95.279	95.5651	95.422	95.1359
	YALE					
Naïve Bayes	53.604	53.604	–	–	–	–
DTNB	60.3977	60.3977	–	–	–	–
OneR	61.599	61.599	–	–	–	–
NNge	–	–	–	–	–	–
JRIP	65.7001	65.7001	–	–	–	–

in Table 9. Especially, the features selected by Filter-CFS are 4.7 and 7 times as RSMI-FS with two distance function. For datasets of SPECT, Heart, and Breast Tumor, the performance of RSMI-FS is better or comparable with others. Because of computational cost of memory, the fuzzy rough methods, Relief, and Filter-CFS can not finish the process of FS about YALE.

Table 9 Numbers of features found by RSMI-FS and other methods

Dataset	FS method					
	RSMI-FS (Euclidean distance)	RSMI-FS (Mahalanobis distance)	VQRS	Fuzzy decision reducts	Relief	Filter-CFS
Small soybean	2	2	2	2	2	8
SPECT	2	2	1	20	2	9
SPECTF	3	2	6	5	3	14
Heart	6	6	7	9	6	3
Breast tumor	4	6	6	4	4	9
YALE	6	6	–	–	–	–

5 An application of RSMI-FS in Alzheimer's disease

Alzheimer's disease (AD) was first identified more than 100 years ago, but the research about its symptoms, causes, risk factors, and treatment has gained momentum only in the past 30 years. Until now, the precise physiological changes triggering the development of AD largely remain unknown (Alzheimer's Association 2012). In this section, we employ machine learning method to find the characteristics of AD.

In order to provide the data of current research initiatives, National Alzheimer's Coordinating Center of US develops a dataset which includes standardized clinical and cognitive data. The dataset is not hypothesis-driven, since all clinical data are developed according to uniform assessment and diagnosed by all participants in AD center. So the research of this dataset is useful to mine the factors triggering AD.

RSMI-FS is a suitable tool to find the relevant factors. Rough sets analyze data using human-like fashion (Jensen and Cornelis 2011b) which is uniform with construction of dataset and process of diagnosing AD. Although rough sets based FS methods are filter methods, the concept of rough sets is classification. So they can find the subsets which improve classification accuracy without dramatical addition of computation cost. Our method has the advantage of rough sets and MI, so the subset selected by RSMI-FS is more relevant and less redundancy. That is to say that RSMI-FS can find the most dangerous factors. In order to find more information of AD, the algorithm stops only when the dependency of selected features arrives at maximum. In Section 4, it is verified that difference between Euclidean distance and Mahalanobis distance is not significant. So Euclidean distance is only used.

The dataset has 11,053 samples and 171 features. Each sample expresses the detail of one person. The features are reduced from dataset, whose proportion of missed values is larger than 70 %. After reducing the features, only 60 features remain. The missed values of the remaining features are estimated by Self-Organizing Map. The detail of Self-Organizing Map is in Kohonen (1982).

In the process of mining factors triggering AD, three aspects must be considered:

1. The quality of results obtained by machine learning method.
2. The number of samples in dataset. It should be large enough to mine the credible factors which trigger the development of AD.
3. The role of label distribution. The proportion of samples with disease is 70.52 %, so it must play role in classification accuracy.

Table 10 Results of AD dataset

Experiment sequence	Number of selected features	Naïve Bayes	DTNB	JRIP	NNge	OneR	Average accuracy	Average of the same size	Proportion of samples with AD(%)
100 samples									
1	2	74	74	71	61	71	70.2	66.54	74
2	3	68	71	70	60	70	67.8		71
3	4	74	73	71	64	64	69.2		73
4	3	72	74	69	67	70	70.4		74
5	4	63	69	63	55	54	60.8		69
6	3	63.6	65.7	59.6	55.6	57.6	60.42		66
7	2	76.8	77.8	75.8	66.7	71.7	73.76		78
8	3	64	67	66	63	56	63.2		67
9	3	66	65	67	60	59	63.4		65
10	3	61	69	67	61	73	66.2		69
500 samples									
1	3	67.1	66.7	65.7	65.1	67.1	66.34	67.54	68
2	4	70.8	71.6	70	62.4	68.2	68.6		71.6
3	4	67.9	68.3	66.1	63.1	66.7	66.42		68.2
4	5	66.2	66.8	67	59	59.6	63.72		67
5	6	66.8	66.8	65.4	59.2	65.2	64.68		68
6	4	70	70.2	70	65	66	68.24		70.2
7	3	70.6	71	69.8	69	68	69.68		71
8	5	71.4	72.6	71.4	65.2	69.4	70		72.6
9	3	71	70.8	71	63.2	68.4	68.88		71
10	3	71	70.8	71	63.2	68.4	68.88		67.4
1,000 samples									
1	9	63.2	70.9	69.1	63.9	68.7	67.16	73.98	70.4
2	5	79.4	79.4	79.3	75.9	79.5	78.7		71.2
3	5	72.5	73.1	72.3	64.7	73.2	71.16		73
4	5	78.9	79.3	80.6	75.8	79.3	78.78		71.6
5	4	70.3	70.4	69.5	65.7	67.2	68.62		71.6
6	5	80.1	80.6	79.8	76	79	79.1		71.8
7	5	80.2	80.2	80.5	76.4	80.5	79.56		68.2
8	4	69.6	69.8	68.7	62.8	70.8	68.34		69.8
9	4	70.7	70.6	70	65	72.6	69.78		71.2
10	5	80.1	79.4	79.4	74.5	79.4	78.56		72
2,000 samples									
1	7	69.8	70.8	70.6	–	69.9	70.275	71.46	70.07
2	7	68.2	71.6	70	–	71.9	70.425		71.2
3	7	68.6	69.7	68	–	69.1	68.85		70.4
4	6	71.9	71.9	71.6	–	71	71.6		71.95
5	9	67.2	68.1	67.6	–	68.1	67.75		68.65
6	11	69.9	70.4	69.9	–	69.9	70.025		71.2
7	7	69.5	69.8	69.1	–	67.6	69		68.2
8	6	79.7	80.1	81.3	–	80.1	80.3		71.75
9	7	66	68.5	68	–	68.6	67.775		68.95
10	5	78.6	78.6	78.6	–	78.7	78.625		70

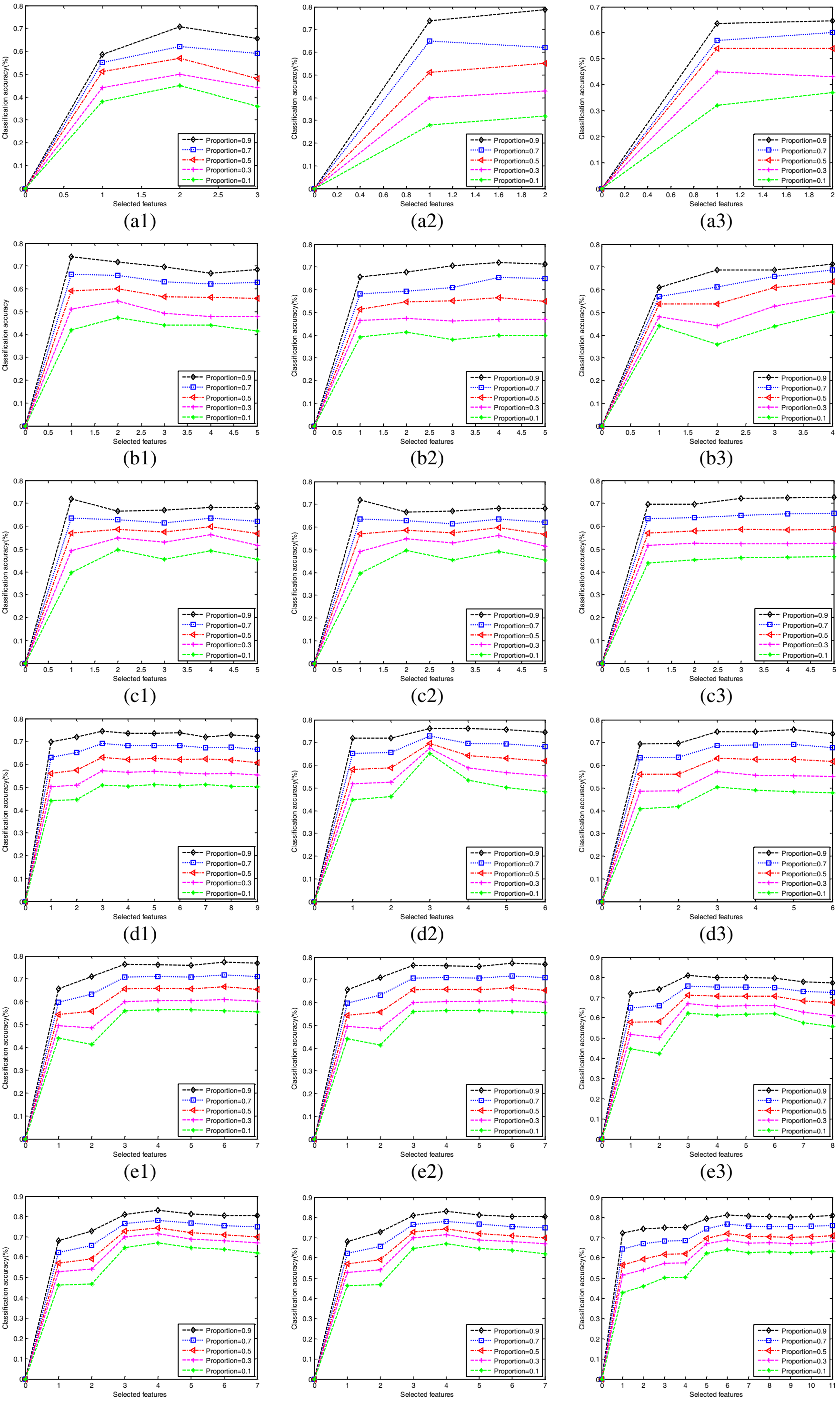
Table 10 (continued)

Experiment sequence	Number of selected features	Naïve Bayes	DTNB	JRIP	NNge	OneR	Average accuracy	Average of the same size	Proportion of samples with AD(%)
3,000 samples									
1	6	79.8	80	80.3	–	79.5	79.9	74.15	70.8
2	9	70.6	71.5	70.9	–	72.1	71.275		71.2
3	7	68.9	69.9	69.1	–	67.9	68.95		69.97
4	7	79.4	80.2	79.9	–	80.2	79.925		71.37
5	7	69.8	71.4	69.7	–	70.5	70.35		70.1
6	8	79.3	79	80.4	–	80	79.675		70.6
7	8	79	80.1	80.7	–	79.3	79.775		70.6
8	8	69.6	69.6	68.9	–	69.4	69.375		70.4
9	7	71.2	70.6	71.2	–	71.3	71.075		71.23
10	7	71.2	71.4	71	–	71.3	71.225		71.6
4,000 samples									
1	8	68.6	70.4	69	–	69.5	69.375	74.64	69.6
2	8	79.1	78.9	80.1	–	79.2	79.325		71.23
3	8	79.1	78.9	80.1	–	79.2	79.325		69.7
4	7	69.6	70.1	70	–	69.9	69.9		70.03
5	11	68.7	70.9	68.5	–	70.6	69.675		69.97
6	6	78.8	80.6	80.6	–	79	79.75		70.6
7	11	69.8	70.2	70.6	–	70.2	70.2		70.3
8	10	68.2	70.1	69.9	–	70.8	69.75		69.9
9	7	79.2	80.7	80.5	–	79.3	79.925		70.45
10	9	78.4	79.3	80	–	78.9	79.15		71.45

The samples should be enough to provide credible results, so the variation of classification accuracies with different amount of samples is evaluated. 100, 500, 1,000, 2,000, 3,000 and 4,000 samples are respectively set as experiment objects. In order to avoid contingency, for each number of samples, 10 different sets are constructed by random sampling. The setup of classification on each subset found by RSMI-FS is the same with Section 4.4. Table 10 shows the details of results, where accuracy is expressed as percentage. For each amount of samples, there are 50 classification accuracies, because of 10 samplings and 5 classifiers. It is hard to directly analyze the variation of classification accuracies with increasing of samples. So the mean of 50 accuracies is used, presented as "Average of the same size" in Table 10. We find the difference among the means of 1,000, 2,000, 3,000, and 4,000 just ranges from 0.484 % to 3.175 %. This guarantees that the samples in the dataset are sufficient to provide precise information. In each row of Table 10, the average accuracy is the average of each sampling set about Naïve Bayes, DTNB, JRIP, NNge

Fig. 8 Variation of classification accuracy under proportion of AD. (**a1**) the training set is one with smallest average accuracy among experiments of 100 samples; (**a2**) the training set is one with largest average accuracy among experiments of 100 samples; (**a3**) the training set is picked randomly among other 8 experiments of 100 samples. (**b1**)–(**b3**) the results of 500 samples are shown in the same way of 100 samples; (**c1**)–(**c3**) 1,000 samples (**d1**)–(**d3**) 2,000 samples. (**e1**)–(**e3**) 3,000 samples. (**f1**)–(**f3**) 4,000 samples

▶



(a1) (a2) (a3)

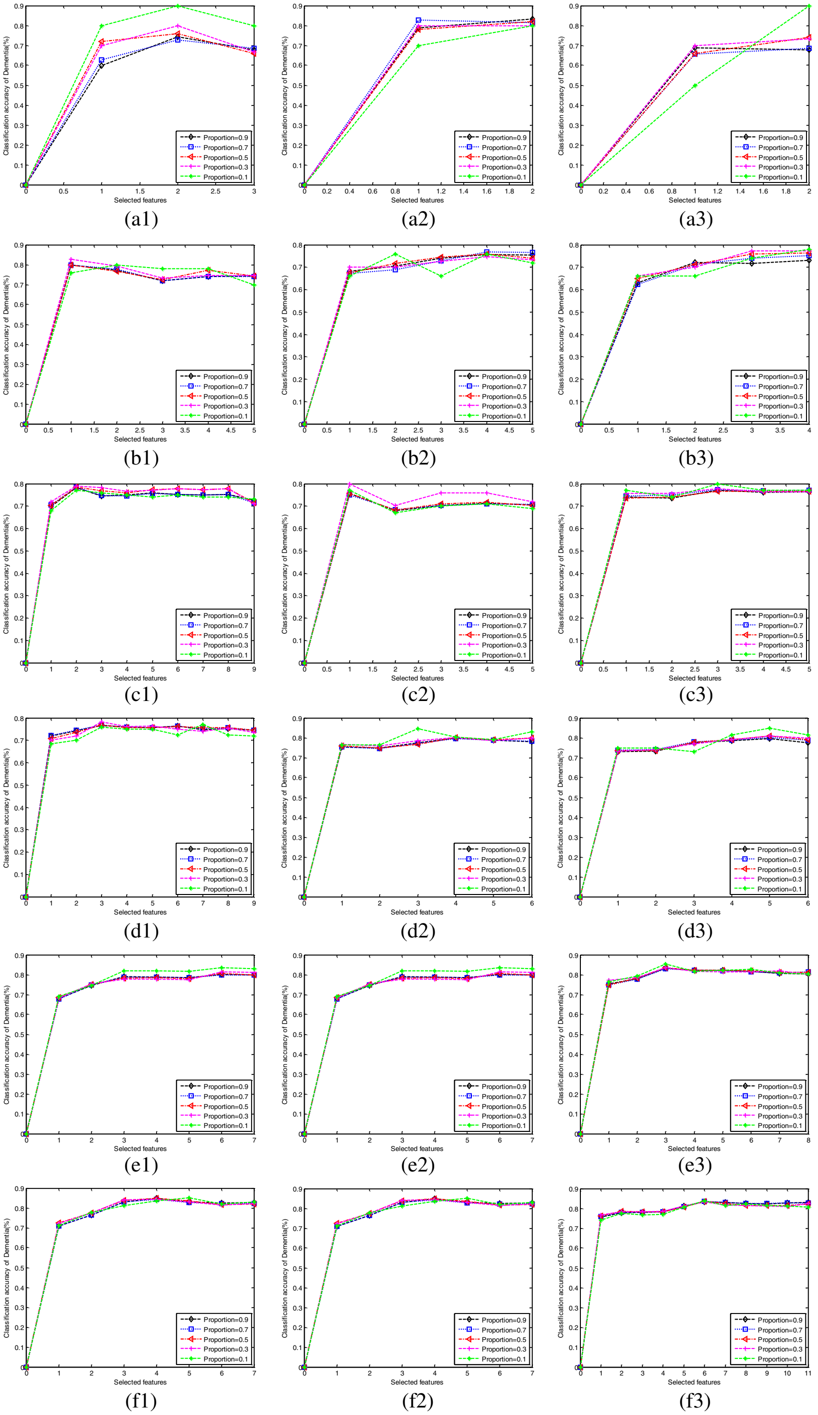
(b1) (b2) (b3)

(c1) (c2) (c3)

(d1) (d2) (d3)

(e1) (e2) (e3)

(f1) (f2) (f3)



(a1) (a2) (a3)

(b1) (b2) (b3)

(c1) (c2) (c3)

(d1) (d2) (d3)

(e1) (e2) (e3)

(f1) (f2) (f3)

Fig. 9 Classification accuracy of samples with diseases. The setup is the same with Fig. 8

and OneR. Average accuracy favors performance evaluation of each sampling set, which is used in following experiment step.

The proportion of samples with AD in the dataset is 70.52 %. It is more dangerous that a person with AD is diagnosed as health, so the center mainly focuses on the people with the disease. But the results of FS should provide service to all kinds of people. That is to say that the proportion of AD in tested people is unknown. It is much possible that healthy people are more than the people with the disease. In order to evaluate effect of the testing sets with different proportion of AD, the setup of new classification is as follows: choose a selected subset from above results as training set; then, a testing set with the same number of samples is formed of samples except in the training set. It must be noted that the proportion of samples with AD in testing set is fixed before random sampling. For each training set, we have 5 testing sets with 0.9, 0.7, 0.5, 0.3, and 0.1 proportion of AD, respectively. The process of choosing training sets: for each number of samples, the selected subset with largest average accuracy and the smallest are both chosen as training sets; one of the 8 others is randomly picked as the third training set. Obviously, all the training sets are the results of FS, so the features in a testing set are reduced according to related training set before classification. A direct classifier, 1NN, is employed in Fig. 8 to avoid role of classifiers. In Fig. 8, the proportion of samples with AD plays significant role in accuracy. With addition of the proportion, the accuracies increase. This result comes to an agreement of label distribution between training and testing set. It must be noted that, for AD dataset, the indicative ability of samples with disease is more important. So Fig. 9 shows the classification accuracies of the samples with disease in the testing sets. It can be observed that the variation of accuracies with proportion in Fig. 9 is not significant. According to the analysis above, RSMI-FS provides the quality results.

6 Conclusion

Feature selection aims to find a small feature subset to represent a given high dimensional dataset. In this paper, we propose a new feature selection method (RSMI-FS) which is based on rough sets theory and mutual information. Attributing to the theoretical contribution of CDM-D-D and FSC criteria, the proposed feature selection method offers three desirable properties over the existing rough sets method. First, the feature subset considers both definitive and uncertain relevance with the labels. The definitive relevance is calculated on the samples in the Positive Region, and the uncertain part implies the relevance information provided by the samples in the Boundary Region. This enables RSMI-FS finds a strong indicative subset of datasets. Second, the proposed method also hybridizes the concept of rough sets with mutual information so that the redundancies of features are reduced by using the FSC. Third, using the dependency of each feature as heuristic information, the results of RSMI-FS are optimal or close-optimal feature subsets. And more importantly, the running time of our proposed RSMI-FS is comparable to the other rough sets based methods.

Acknowledgement National Alzheimer's Coordinating Center, funded by the National Institute on Aging (UO1 AG016976), provides the uniform Alzheimer's disease data set (2005–2011).

References

Aha, D. (1992). Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies, 36*(2), 267–287.

Alzheimer's Association (2012). 2012 Alzheimer's disease facts and figures. *Alzheimer's & Dementia, 8*, 131–168.

Aouiche, K., & Darmont, J. (2009). Data mining-based materialized view and index selection in data warehouses. *Journal of Intelligent Information Systems, 33*, 65–93.

Bae, C., Yeh, W., Chun, Y., Liu, S. (2010). Feature selection with intelligent dynamic swarm and rough set. *Expert Systems with Applications, 37*(10), 7026–7032.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Network, 5*(4), 537–550.

Bazan, J., Nguyen, H., Nguyen, S., Synak, P., Wróblewski, J. (2000). Rough set algorithms in classification problem. In *Proc. of rough set methods and applications* (pp. 49–88). Germany: Physica-Verlag Gmbh Heidelberg.

Bonnlander, B. (1996). *Nonparametric selection of input variables for connectionist learning*. Ph.D. Thesis, CU-CS-812-96, University of Colorado, Boulder.

Breiman, L., Friedman, J., Stoneand, C., Olshen, R. (1984). *Classification and regression trees*. Boca Raton: CRC Press.

Chen, Y., Miao, D., Wang, R., Wu, K. (2011). A rough set approach to feature selection based on power set tree. *Knowledge-Based Systems, 24*(2), 275–281.

Chow, T., & Huang, D. (2005). Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information. *IEEE Transactions on Neural Networks, 16*(1), 213–223.

Cohen, W. (1995). Fast effective rule induction. In *Proc. of 12th international conference on machine learning* (pp. 115–123). Lake Taho, CA.

Cornelis, C., & Jensen, R. (2008). A noise-tolerant approach to fuzzy-rough feature selection. In *Proc. of IEEE world congress on computational intelligence (FUZZ-IEEE 2008)* (pp. 1598–1605). Hong Kong, China.

Cornelis, C., Martín, G., Jensen, R., Ślęzak, D. (2008). Feature selection with fuzzy decision reducts. In *Proc. of 3th international conference on rough sets and knowledge technology (RSKT 2008)* (pp. 284–291). Chengdu, China.

Cornelis, C., Jensen, R., Martín, G., Ślęzak, D. (2010a). Attribute selection with fuzzy decision reducts. *The Information of the Science, 180*(2), 209–224.

Cornelis, C., Verbiest, N., Jensen, R. (2010b). Ordered weighted average based fuzzy rough sets. In *Proc. of 5th international conference on Rough Sets and Knowledge Technology (RSKT 2010)* (pp. 78–85). Beijing, China.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York: Wiley.

Dash, M., & Liu, H. (2003). Consistency-based search in feature selection. *Artificial Intelligence, 151*(1–2), 155–176.

Delimata, P., & Suraj, Z. (2008). Feature selection algorithm for multiple classifier systems: a hybrid approach. *Fundamenta Informaticae, 85*(1–4), 97–110.

Delimata, P., Moshkov, M., Skowron, A., Suraj, Z. (2008). Comparison of lazy classification algorithms based on deterministic and inhibitory decision rules. In *Proc. of 3th international conference on rough sets and knowledge technology (RSKT 2008)* (pp. 17–19). Chengdu, China; Lecture Notes in Artificial Intelligence, 5009/2008: 55–62.

Delimata, P., Moshkov, M., Skowron, A., Suraj, Z. (2009). *Inhibitory rules in data analysis: a rough set approach. Studies in computational intelligence* (Vol. 63). Springer-Verlag, Berlin.

Delimata, P., Marszał-Paszek, B., Moshkov, M., Paszek, P., Skowron, A., Suraj, Z. (2010). Comparison of some classification algorithms based on deterministic and nondeterministic decision rules. Transactions on Rough Sets XII. *Lecture Notes of Computer Science, 6190*(2010), 90–105.

ElAlami, M. (2009). A filter model for feature subset selection based on genetic algorithm. *Knowledge-Based Systems, 22*(5), 356–362.

Estévez, P., Tesmer, M., Perez, C., Zurada, J. (2009). Normalized mutual information feature selection. *IEEE Transactions on Neural Networks, 20*(2), 189–201.

Georghiades, A., Belhumeur, P., Kriegman, D. (2001). From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(6), 643–660.

Grzymala-Busse, J., & Rzasa, W. (2006). Local and global approximations for incomplete data. Rough sets and current trends in computing. *Lecture Notes in Computer Science, 4259*(2006), 244–253.

Guetlein, M., Frank, E., Hall, M., Karwath, A. (2009). Large scale attribute selection using wrappers. In *Proc of IEEE symposium on computational intelligence and data mining* (pp. 332–339).

Hall, M. (1998). *Correlation-based feature subset selection for machine learning*. Ph.D. Thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand.

Hall, M., & Frank, E. (2008). Combining naive Bayes and decision tables. In *Proc. of 21st florida artificial intelligence research society conference* (pp. 318–319). Miami, Florida.

Holte, R. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning, 11*(1), 63–90.

Hu, Q., Xie, Z., Yu, D. (2007). Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation. *Pattern Recognition, 40*(12), 3509–3521.

Hu, Q., Che, X., Zhang, L., Yu, D. (2010). Feature evaluation and selection based on neighborhood soft margin. *Neurocomputing, 73*(10–12), 2114–2124.

Huang, D., & Chow, T. (2005). Effective feature selection scheme using mutual information. *Neurocomputing, 63*, 325–343.

Jensen, R., & Cornelis, C. (2011a). Fuzzy-rough nearest neighbour classification. Transactions on rough sets XIII. *Lecture Notes in Computer Science, 6499*(2011), 56–72.

Jensen, R., & Cornelis, C. (2011b). Fuzzy-rough nearest neighbour classification and prediction. *Theoretical Computer Science, 412*(42), 5871–5884.

Jensen, R., Cornelis, C., Shen, Q. (2009). Hybrid fuzzy-rough rule induction and feature selection. In *Proc. of IEEE World congress on computational intelligence (FUZZ-IEEE 2009)* (pp. 1151–1156). Jeju Island, Korea.

John, G., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Proc. of 17th conference on uncertainty in artificial intelligence* (pp. 338–345). San Mateo: Morgan Kaufmann.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*, 59–69.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of Relief. In *European conference on machine learning* (pp. 171–182).

Kumar, A. (1998). New techniques for data reduction in a database system for knowledge discovery applications. *Journal of Intelligent Information Systems, 10*, 31–48.

Kuncheva, L., & Jain, L. (1999). Nearest neighbor classifier: simultaneous editing and feature selection. *Pattern Recognition Letters, 20*(11–13), 1149–1156.

Kwak, N., & Choi, C. (2002a). Input feature selection for classification problems. *IEEE Transactions on Neural Networks, 13*(1), 143–159.

Kwak, N., & Choi, C. (2002b). Input feature selection by mutual information based on parzen window. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(12), 1667–1671.

Liu, H., Sun, J., Liu, L., Zhang, H. (2009). Feature selection with dynamic mutual information. *Pattern Recognition, 42*(7), 1330–1339.

Martin, B. (1995). *Instance-based learning: Nearest neighbour with generalization*. Master Thesis, University of Waikato, Hamilton, New Zealand.

Moshkov, M., Skowron, A., Suraj, Z. (2008). Extracting relevant information about reduct sets from data tables. Transactions on rough sets IX. *Lecture Notes of Computer Science, 5390*(2008), 200–211.

Moshkov, M., Skowron, A., Suraj, Z. (2010). Irreducible descriptive sets of attributes for information systems. Transactions on rough sets XI. *Lecture Notes of Computer Science, 5964*(2010), 92–105.

Parthaláin, N., & Shen, Q. (2009). Exploring the boundary region of tolerance rough sets for feature selection. *Pattern Recognition, 42*(5), 655–667.

Parthaláin, N., Shen, Q., Jensen, R. (2007). Distance measure assisted rough set feature selection. In *Proc. of 16th international conference on fuzzy systems* (pp. 1084–1089).

Parthaláin, N., Shen, Q., Jensen, R. (2010). A distance measure approach to exploring the rough set boundary region for attribute reduction. *IEEE Transactions on Knowledge and Data Engineering, 22*(3), 305–317.

Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Science, 11*(5), 341–356.

Pawlak, Z. (1991). *Rough sets—theoretical aspects of reasoning about data* (pp. 33–44). London: Kluwer Academic Publishers.

Pawlak, Z., & Skowron, A. (2007a). Rudiments of rough sets. *Information Sciences, 177*(1), 3–27.

Pawlak, Z., & Skowron, A. (2007b). Rough sets: some extensions. *Information Sciences, 177*(1), 28–40.

Pawlak, Z., & Skowron, A. (2007c). Rough sets and boolean reasoning. *Information Sciences, 177*(1), 41–73.

Peng, H., Long, F., Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(8), 1226–1238.

Quinlan, J. (1993). *C4.5: Programs for machine learning*. San Mateo: Morgan Kaufmann Publishers.

Robnik-Sikonja, M., & Kononenko, I. (1997). An adaptation of relief for attribute estimation in regression. In *14th international conference on machine learning* (pp. 296–304).

Ron, K., & John, G. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1–2), 273–324.

Słowiński, R. (1992). *Intelligent decision support: Handbook of application and advances of the rough sets theory*. London: Kluwer Academic Publishers.

Słowiński, R., & Stefanowski, J. (1989). Rough classification in incomplete information systems. *Mathematical and Computer Modelling, 12*(10–11), 1347–1357.

Sotoca, J., & Pla, F. (2010). Supervised feature selection by clustering using conditional mutual information-based distances. *Pattern Recognition, 43*(6), 2068–2081.

Stefanowski, J., & Tsoukiàs, A. (2001). Incomplete information tables and rough classification. *Computational Intelligence, 17*(3), 545–566.

Torkkola, K., & Campbell, W. (2000). Mutual information in learning feature transformations. In *Proc. of 17th international conference on machine learning* (pp. 1015–1022). Stanford, CA, USA.

Wang, X., Yang, J., Teng, X., Xia, W., Jensen, R. (2007). Feature selection based on rough sets and particle swarm optimization. *Pattern Recognition Letters, 28*(4), 459–471.

Xu, Z., King, I., Lyu, M., Rong, J. (2010). Discriminative semi-supervised feature selection via manifold regularization. *IEEE Transactions on Neural Networks, 21*(7), 1033–1047.

Yao, Y., Zhao, Y., Wang, J., Han, S. (2006). A model of machine learning based on user preference of attributes. In *Proc. of 5th international conference on rough sets and current trends in computing*. *Lecture Notes of Computer Science* (Vol. 4259, pp. 587–596).

Yao, Y., Zhao, Y., Wang, J. (2008). On reduct construction algorithms. *LNCS Transactions on Computational Science II, LNCS, 5150*, 100–117.

Zhong, N., Dong, J., Ohsuga, S. (2001). Using rough sets with heuristics for feature selection. *Journal of Intelligent Information Systems, 16*, 199–214.