Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 989067, 10 pages
http://dx.doi.org/10.1155/2015/989067

Research Article

A Mixed Feature Selection Method Considering Interaction

Zilin Zeng,[1,2] Hongjun Zhang,[1] Rui Zhang,[1] and Youliang Zhang[1]

[1] *PLA University of Science & Technology, Nanjing 210007, China*
[2] *Nanchang Military Academy, Nanchang 330103, China*

Correspondence should be addressed to Zilin Zeng; beauty1981@sohu.com

Received 26 May 2014; Revised 18 November 2014; Accepted 22 November 2014

Academic Editor: Ivanka Stamova

Feature interaction has gained considerable attention recently. However, many feature selection methods considering interaction are only designed for categorical features. This paper proposes a mixed feature selection algorithm based on neighborhood rough sets that can be used to search for interacting features. In this paper, feature relevance, feature redundancy, and feature interaction are defined in the framework of neighborhood rough sets, the neighborhood interaction weight factor reflecting whether a feature is redundant or interactive is proposed, and a neighborhood interaction weight based feature selection algorithm (NIWFS) is brought forward. To evaluate the performance of the proposed algorithm, we compare NIWFS with other three feature selection algorithms, including INTERACT, NRS, and NMI, in terms of the classification accuracies and the number of selected features with C4.5 and IB1. The results from ten real world datasets indicate that NIWFS not only deals with mixed datasets directly, but also reduces the dimensionality of feature space with the highest average accuracies.

1. Introduction

Feature selection plays an important role in pattern recognition and machine learning. It has drawn attention of many researchers from various fields. This task aims to select the essential features that allow us to discern between patterns belonging to different classes. The effects of feature selection have been widely recognized in, for example, improving predictive accuracy, facilitating data visualization, reducing storage requirements, and reducing training time [1].

Many feature selection methods have been proposed to remove as many irrelevant and redundant features as possible [2–8], such as Relief and its variation Relief-F [9, 10], correlation-based feature selection (CFS) [11], mutual information based feature selection (MIFS) [12], fast correlation-based filter (FCBF) [13], and minimum-redundancy maximum-relevance (MRMR) [14]. However, apart from the identification of irrelevant and redundant features, an important but usually neglected issue is feature interaction [15]. Interacting features are those that appear to be irrelevant or weakly relevant to the class individually, but when it is combined with other features, it may highly correlate to the class. A typical example is the XOR problem. There are two features and a class label which is zero if both features have the same value and one otherwise. Obviously, each feature does not carry any information about the class individually; however, the two features determine the class completely when combined. In many classification problems, a feature that is completely useless by itself sometimes can provide a significant performance improvement when taken with others. If we only consider relevance and redundancy but ignore interaction, some salient features may be missing.

Some wrapper methods are able to deal with feature interaction to some extent, but these methods require a model testing each feature subset and the process is usually time-consuming, especially for some computational expensive models. Furthermore, wrapper methods are very sensitive to the specific classification algorithm, and the performance of the model does not necessarily reflect the actual predictive ability of the selected feature subset. Therefore, it is a challenge to filter out the irrelevant and redundant features and reserve only a small number of interactive features. Feature interaction increasingly arouses the attention of researchers. Zhao and Liu [16] propose to search for interacted features using consistency contribution to measure feature relevance. Recently, Wang et al. [17] bring forward a propositional

FOIL rule based algorithm FRFS. The algorithm involves two steps: (i) redundant feature exclusion and interactive feature reservation and (ii) the irrelevant feature identification. The experimental results demonstrate the effectiveness of FRFS algorithm.

Although the work mentioned above has pointed out the existence and effectiveness of feature interaction, the state-of-the-art feature selection techniques of searching interacting features are merely designed for categorical datasets. In real world, however, data comes with a mixed format in the majority of cases [18]. Discretizing numerical features usually bring information loss because the degrees of membership of values to discretized values are not considered [19].

Rough set theory [20–23], introduced by Pawlak, is a well-known mathematical approach addressing vague and uncertain data with no additional information. It has attracted the attention of many researchers who have studied its theories and applications during the last decades. Rough set theory can be used to find a subset of informative features which preserves the discernible ability from the original features. Therefore, it has been playing an important role in feature selection [24–29]. However, classical rough set theory can only deal with nominal feature values. Since numerical feature values are more common in real world, the crisp rough set theory encounters a challenge. Therefore, some new models such as fuzzy rough sets and neighborhood rough sets are usually considered for extension of the classical rough set theory. These extended models can be used to deal with mixed numerical and categorical data within a uniform framework [30–32]. For example, Jensen and Shen [33, 34] generalized the dependency function defined in classical rough sets based on positive region into the fuzzy case and presented a rough-fuzzy feature selection algorithm. Hu et al. [31, 35] substituted classical equivalence relation with neighborhood relations and introduced a neighborhood rough sets model to address the data with mixed features. A neighborhood rough set based heterogeneous feature subset selection (NRS) [31] which utilizes the neighborhood dependency to evaluate the significance of a subset of heterogeneous features is proposed. As the robustness to noise and transformation of mutual information, Hu et al. also generalized Shannon's information entropy to neighborhood information entropy and proposed a neighborhood mutual information based feature selection method (NMI) [35].

Inspired by the fact that neighborhood granules can characterize numerical features, in this paper, we attempt to analyze relevance, redundancy, and interaction in the framework of neighborhood rough sets. Since redundant features produce negative influence and interaction features produce positive influence in predicting, a neighborhood interaction weight factor is introduced to measure the redundancy and interaction of candidate features. We can adjust the relevance measure between a feature and the class by the neighborhood interaction weight factor and rank the candidate features with the adjusted relevance measure. Finally, we propose a neighborhood interaction weight based feature selection algorithm (NIWFS). To verify its performance, the proposed method is compared with three state-of-the-art feature selection methods (INTERACT, NRS, and NMI) on a series of benchmark datasets. Experiment results show that our proposed method can be applied to dataset with mixed categorical and numerical features directly and outperforms the other selectors.

The remainder of this paper is structured as follows: Section 2 reviews some basic concepts related to neighborhood rough sets and neighborhood entropy-based information measures; Section 3 provides our definitions of relevant feature, redundant feature, and interactive feature based on neighborhood interaction gain; Section 4 puts forward a neighborhood interaction weight based feature selection algorithm; Section 5 presents the experimental results and analysis to evaluate the effectiveness of the proposed method; and Section 6 lays out our conclusions.

2. Preliminaries

In this section, we briefly introduce some basic concepts and notations of neighborhood rough set model and some neighborhood entropy-based information measures.

2.1. Neighborhood Rough Set Model. The notion of an information system provides a convenient basis for the representation of objects in terms of their attributes (also called features). An information system is a quadruple (U, A, V, f), where U is a nonempty finite set of objects called the universe, A is a nonempty finite set of attributes, $V = \bigcup_{a \in A} V_a$ where V_a is the value domain of attribute a, and $f : U \times A \to V$ is an information function which associates a unique value of each attribute with every object belonging to U such that, for any $a \in A$ and $u \in U$, $f(u, a) \in V_a$. A decision table $\mathrm{IS} = (U, C \cup \{d\}, V, f)$ is a special case of information system, where attributes in A are called condition attributes and d is a designated attribute called the decision attribute.

Definition 1 (see [36]). A neighborhood information system is a quintuple $\mathrm{NIS} = (U, A, V, f, \delta)$, where U is a nonempty finite set of objects called the universe; A is a nonempty finite set of attributes; V is the union of attribute domains such that $V = \bigcup_{a \in A} V_a$; for any $a \in A$, there exists a mapping $U \to V_a$, where V_a is the set of values of a; δ is a neighborhood parameter. More specially, $\mathrm{NIS} = (U, C \cup \{d\}, V, f, \delta)$ is called a neighborhood decision system, where C is a set of condition attributes and d is a decision attribute.

In classical rough sets, the objects with the same feature value are pooled into a set, called equivalence class. These objects are expected to belong to the same class; otherwise, they are inconsistent. However, it is unfeasible to compute equivalence classes with numerical features because the probability of objects with the same numerical value is very small [35]. Therefore, the equivalence class will be substituted by the neighborhood class.

Definition 2 (see [36]). Let $\mathrm{NIS} = (U, A, V, f, \delta)$ be a neighborhood information system. For any attribute subset $B \subseteq A$, there exists a distance function $d_B : U \times U \to [0, 1]$, and then B determines a similarity relation denoted by $\mathrm{NR}_\delta(B)$ as follows:

$$\mathrm{NR}_\delta (B) = \{(x, y) \in U \times U \mid d_B (x, y) \leq \delta\} . \tag{1}$$

The neighborhood class $N_B^\delta(x)$ with respect to B is defined as

$$N_B^\delta(x) = \{y \mid y \in U, d(x,y) \le \delta\}. \tag{2}$$

The distance metric can be denoted by

$$d_B^P(x,y) = \left(\sum_{i=1}^{N} \left|a_i(x) - a_i(y)\right|^P\right)^{1/P}. \tag{3}$$

This is a Manhattan distance if $P = 1$, a Euclidean distance if $P = 2$, and a Chebyshev distance if $P = \infty$.

Theorem 3 (monotonicity [31]). *Let $NIS = (U, A, V, f, \delta)$ be a neighborhood information system. For $P, Q \subseteq A$ and $x_i \in U$, one has the following:*

(1) *if $P \subseteq Q$, then $N_P^\delta(x_i) \supseteq N_Q^\delta(x_i)$;*

(2) *if $\delta_1 \le \delta_2$, then $N_P^{\delta_1}(x_i) \subseteq N_P^{\delta_2}(x_i)$.*

The monotonicity is very important for constructing a greedy forward or backward search algorithm [2]. It guarantees that addition of any new attribute to the existing subset does not lead to a decrease of the relevance between the new subset and the decision attribute.

2.2. Some Neighborhood Entropy-Based Information Measures. Shannon's information theory, first introduced in 1948 [37], provides a way to measure the information of random variables. The entropy is a measure of uncertainty of random variables [38]. In this section, some neighborhood entropy-based information measures are defined in a neighborhood system which is the generalization of Shannon's entropy.

Definition 4 (see [35]). Let NIS $= (U, A, V, f, \delta)$ be a neighborhood information system, $B \subseteq A$, and $x_i \in U$. Then the neighborhood entropy is defined as

$$\mathrm{NH}_\delta(B) = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log \frac{\left|N_B^\delta(x_i)\right|}{|U|}. \tag{4}$$

Since $\forall x_i$, $N_B^\delta \subseteq U$, $1/|U| \le |N_B^\delta(x_i)|/|U| \le 1$, so we have $0 \le \mathrm{NH}_\delta(B) \le \log|U|$. $\mathrm{NH}_\delta(B) = \log|U|$ if and only if $\forall x_i$, $|N_B^\delta(x_i)| = 1$; that is, $N_B^\delta(x_i) = \{x_i\}$. $\mathrm{NH}_\delta(B) = 0$ if and only if $\forall x_i$, $|N_B^\delta(x_i)| = |U|$; that is, $N_B^\delta(x_i) = U$. Obviously, when attribute subset B can distinguish any two objects, the neighborhood entropy is the largest; when attribute subset B cannot distinguish any two objects, the neighborhood entropy is zero.

Theorem 5 (see [35]). *If $\delta = 0$, then $NH_\delta(B) = H(B)$, where $H(B)$ is Shannon's entropy.*

Theorem 5 indicates that neighborhood entropy equals Shannon's entropy if attributes are discrete. That is, neighborhood entropy is a natural generalization of Shannon's entropy.

Definition 6 (see [35]). Let NIS $= (U, A, V, f, \delta)$ be a neighborhood information system, $P, Q \subseteq A$, and $x_i \in U$. Then the joint neighborhood entropy is computed as

$$\mathrm{NH}_\delta(P,Q) = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log \frac{\left|N_{P\cup Q}^\delta(x_i)\right|}{|U|}. \tag{5}$$

Theorem 7 (see [35]). *Consider $NH_\delta(P,Q) \ge NH_\delta(P)$, $NH_\delta(P,Q) \ge NH_\delta(Q)$.*

Definition 8 (see [35]). Let NIS $= (U, A, V, f, \delta)$ be a neighborhood information system, $P, Q \subseteq A$, and $x_i \in U$. Then the neighborhood mutual information of P and Q is defined as

$$\mathrm{NMI}_\delta(P;Q) = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log \frac{\left|N_P^\delta(x_i)\right| \cdot \left|N_Q^\delta(x_i)\right|}{|U| \cdot \left|N_{P\cup Q}^\delta(x_i)\right|}. \tag{6}$$

The neighborhood mutual information $\mathrm{NMI}_\delta(P;Q)$ describes the common information found in P and Q. It is usually used to measure the relevance between numerical or nominal variables.

Theorem 9 (see [35]). *The relation between neighborhood mutual information and neighborhood entropy is as follows:*

(1) *$NMI_\delta(P;Q) = NMI_\delta(Q;P)$;*

(2) *$NMI_\delta(P;Q) = NH_\delta(P) + NH_\delta(Q) - NH_\delta(P,Q)$;*

(3) *$NMI_\delta(P;Q) = NH_\delta(P) - NH_\delta(P \mid Q) = NH_\delta(Q) - NH_\delta(Q \mid P)$.*

Definition 10. Let NIS $= (U, A, V, f, \delta)$ be a neighborhood information system, $P, Q, R \subseteq A$, and $x_i \in U$. Then the conditional neighborhood mutual information of P and Q given R is defined as

$$\mathrm{NMI}_\delta(P;Q \mid R) = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log \frac{\left|N_{P\cup R}^\delta(x_i)\right| \cdot \left|N_{P\cup Q}^\delta(x_i)\right|}{\left|N_{P\cup Q\cup R}^\delta(x_i)\right| \cdot \left|N_P^\delta(x_i)\right|}. \tag{7}$$

Theorem 11. *Consider $NMI_\delta(P;Q \mid R) = NMI_\delta(P,R;Q) - NMI_\delta(P;Q)$.*

Proof. Consider the following:

$$\begin{aligned}
&\mathrm{NMI}_\delta(P,R;Q) - \mathrm{NMI}_\delta(P;Q)\\
&\quad = \mathrm{NMI}_\delta(P \cup R;Q) - \mathrm{NMI}_\delta(P;Q)\\
&\quad = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log \frac{\left|N_{P\cup R}^\delta(x_i)\right| \cdot \left|N_Q^\delta(x_i)\right|}{|U| \cdot \left|N_{P\cup R\cup Q}^\delta(x_i)\right|}\\
&\qquad - \left(-\frac{1}{|U|}\sum_{i=1}^{|U|} \log \frac{\left|N_P^\delta(x_i)\right| \cdot \left|N_Q^\delta(x_i)\right|}{|U| \cdot \left|N_{P\cup Q}^\delta(x_i)\right|}\right)
\end{aligned}$$

$$\begin{aligned}
&= -\frac{1}{|U|}\sum_{i=1}^{|U|}\left(\log\frac{\left|N^{\delta}_{P\cup R}(x_i)\right|\cdot\left|N^{\delta}_{Q}(x_i)\right|}{|U|\cdot\left|N^{\delta}_{P\cup R\cup Q}(x_i)\right|}\right.\\
&\qquad\left.-\log\frac{\left|N^{\delta}_{P}(x_i)\right|\cdot\left|N^{\delta}_{Q}(x_i)\right|}{|U|\cdot\left|N^{\delta}_{P\cup Q}(x_i)\right|}\right)\\
&= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{\left|N^{\delta}_{P\cup R}(x_i)\right|\cdot\left|N^{\delta}_{P\cup Q}(x_i)\right|}{\left|N^{\delta}_{P\cup R\cup Q}(x_i)\right|\cdot\left|N^{\delta}_{P}(x_i)\right|}\\
&= \mathrm{NMI}_{\delta}(P;Q\mid R).
\end{aligned} \tag{8}$$

□

Theorem 11 shows that the conditional neighborhood mutual information is the reduction in the uncertainty of P due to knowledge of Q when R is given.

3. Relevance, Redundancy, and Interaction

The concepts such as relevance, redundancy, and interaction of features have been used frequently in the study of feature selection. However, a quantitative formalism for mixed data has not been available to date. In this section, we will redefine the relevant feature, redundant feature, and interactive feature by using neighborhood information measures.

In the discrete case, mutual information has been frequently used to evaluate the strength of the relevance between a feature F_i and the class C. In this situation, the features are evaluated individually. However, some features influence the class variable by grouping rather than individualizing. A well-known illustration of this phenomenon is the XOR problem as shown in Table 1.

Table 1: XOR problem.

F_1	F_2	$C = F_1 \oplus F_2$
1	0	1
0	1	1
1	1	0
0	0	0

One can see that F_1 and F_2 have a null relevance individually; that is, $\mathrm{MI}(F_1;C) = 0$, $\mathrm{MI}(F_2;C) = 0$. That is to say, feature F_1 (or F_2) can be considered irrelevant in terms of mutual information. However, when we combine F_1 and F_2, the maximal relevance is obtained; that is, $\mathrm{MI}(F_1,F_2;C) = H(C) = 1$. This indicates that feature F_1 (or F_2) is strongly relevant to the class C. Moreover, the mutual information is difficult to compute when the features are continuous. Therefore, it is necessary to redefine the relevant feature. We give the new definition of the relevant feature as follows.

Definition 12 (relevant feature). Let $\mathbb{F}$ be a full set of features, $F_i \in \mathbb{F}$, and $F' = \mathbb{F} - \{F_i\}$. Feature F_i is relevant to the class label C if and only if

$\exists S \subseteq F'$, such that $\mathrm{NMI}_{\delta}(F_i;C\mid S) > 0$.

Otherwise, F_i is an irrelevant feature.

According to Definition 12, relevance should be conditionally dependent on the context S. It is easy to find that $\mathrm{NMI}_{\delta}(F_1;C\mid F_2) = \mathrm{MI}(F_1;C\mid F_2) > 0$ and $\mathrm{NMI}_{\delta}(F_2;C\mid F_1) = \mathrm{MI}(F_2;C\mid F_1) > 0$ in Table 1. Hence, features F_1 and F_2 have become relevant features under the new definition.

Previous work mostly focuses on the definitions of relevant features and redundant features [39]. Interactive features are often ignored. To judge whether there exists interaction or redundancy between features, we introduce the neighborhood interaction gain by using neighborhood information measures.

Definition 13 (neighborhood interaction gain). Let $\mathbb{F}$ be a full set of features, $F_i, F_j \in \mathbb{F}$ and the class C; then the neighborhood interaction gain is defined as

$$\begin{aligned}
\mathrm{NIG}_{\delta}\left(F_i,F_j\right) &= \mathrm{NMI}_{\delta}\left(F_i,F_j;C\right)\\
&\quad - \mathrm{NMI}_{\delta}\left(F_i;C\right) - \mathrm{NMI}_{\delta}\left(F_j;C\right).
\end{aligned} \tag{9}$$

By Theorem 11, we also have

$$\begin{aligned}
\mathrm{NIG}_{\delta}\left(F_i,F_j\right) &= \mathrm{NMI}_{\delta}\left(F_i;C\mid F_j\right) - \mathrm{NMI}_{\delta}\left(F_i;C\right)\\
&= \mathrm{NMI}_{\delta}\left(F_i;C\mid F_j\right) - \mathrm{NMI}_{\delta}\left(F_i;C\right).
\end{aligned} \tag{10}$$

Neighborhood interaction gain can be interpreted as the change in a dependence between feature F_i (or F_j) and the class C by introducing context F_j (or F_i). It is quite easy to see that when the neighborhood interaction gain is negative, context decreases the amount of dependence. When the neighborhood interaction gain is positive, context increases the amount of dependence. When the interaction gain is zero, context does not affect the dependence between feature F_i (or F_j) and the class C.

If the neighborhood interaction gain is positive, we benefit from a synergy between the features F_i and F_j [40, 41]. In other words, the addition of feature F_j will produce positive influence in predicting C for F_i. A well-known example of such synergy is the XOR problem. If the neighborhood interaction gain is negative, we suffer diminishing returns by several features providing overlapping, redundant information. In fact, the neighborhood interaction gain is the amount of information gained (or lost) in transmission by controlling one feature when the other feature is already known [42]. Based on this, we give the new definitions of redundant and interactive feature in the following.

Definition 14 (redundant feature). Letting $\mathbb{F}$ be a full set of features, $F_i, F_j \in \mathbb{F}$, feature F_i is said to be redundant with feature F_j if and only if

$$\mathrm{NIG}_{\delta}\left(F_i,F_j\right) \le 0. \tag{11}$$

According to Definition 14, $\mathrm{NIG}_{\delta}(F_i,F_j) \le 0$ suggests a redundancy between F_i and F_j; in other words, they both provide in part the same information about the class C. Therefore, the inequality implies that F_i is a redundant feature when given feature F_j.

Definition 15 (interactive feature). Letting $\mathbb{F}$ be a full set of features, $F_i, F_j \in \mathbb{F}$, feature F_i is said to be interactive with feature F_j if and only if

$$\mathrm{NIG}_{\delta}\left(F_i, F_j\right) \geq 0. \tag{12}$$

According to Definition 15, $\mathrm{NIG}_{\delta}(F_i, F_j) \geq 0$ indicates a synergy between feature F_i and F_j; that is, they yield more information together than what could be expected from the sum of $\mathrm{NMI}_{\delta}(F_i; C)$ and $\mathrm{NMI}_{\delta}(F_j; C)$. In other words, the absence of either feature will decrease the ability of predicting the class C.

4. Proposed Feature Selection Algorithm

In this section, we define the neighborhood interaction weight factor based on the neighborhood interaction gain and then move on to present our proposed feature subset selection algorithm.

4.1. Neighborhood Interaction Weight Factor. One can see that the introduction of feature F_j affects the dependence between the feature F_i and the class C. The positive neighborhood interaction gain means that we cannot depict their relationship without considering both of them at once and the addition of another feature will increase the amount of dependence. That is to say, the introduction of feature F_j has a positive influence on predicting the class variable C. Correspondingly, we should increase the weight of feature F_j. The negative neighborhood interaction gain means that the introduction of the new feature will inhibit the amount of dependence. That is to say, the introduction of feature F_j has a negative influence in predicting the class variable C. Correspondingly, we should decrease the weight of feature F_j. Therefore, we can define the neighborhood interaction weight factor based on the neighborhood interaction gain. It is possible to analyze relationships between features and guide feature selection and construction.

Definition 16 (neighborhood interaction weight factor). The neighborhood interaction weight factor of feature F_i with respect to feature F_j is defined as

$$\mathrm{NIW}_{\delta}\left(F_i, F_j\right) = 1 + \frac{\mathrm{NIG}_{\delta}\left(F_i, F_j\right)}{\mathrm{NH}_{\delta}\left(F_i\right) + \mathrm{NH}_{\delta}\left(F_j\right)}. \tag{13}$$

Theorem 17. *Consider* $0 \leq NIW_{\delta}(F_i, F_j) \leq 2$.

Proof. Since $0 \leq \mathrm{NMI}_{\delta}(F_i, F_j; C) \leq \mathrm{NH}_{\delta}(F_i, F_j) \leq \mathrm{NH}_{\delta}(F_i) + \mathrm{NH}_{\delta}(F_j)$, $0 \leq \mathrm{NMI}_{\delta}(F_i; C) \leq \mathrm{NH}_{\delta}(F_i)$, $0 \leq \mathrm{NMI}_{\delta}(F_j; C) \leq \mathrm{NH}_{\delta}(F_j)$, and $\mathrm{NIG}_{\delta}(F_i, F_j) = \mathrm{NMI}_{\delta}(F_i, F_j; C) - \mathrm{NMI}_{\delta}(F_i; C) - \mathrm{NMI}_{\delta}(F_j; C)$, we have $-[\mathrm{NH}_{\delta}(F_i) + \mathrm{NH}_{\delta}(F_j)] \leq \mathrm{NIG}_{\delta}(F_i, F_j) \leq \mathrm{NH}_{\delta}(F_i) + \mathrm{NH}_{\delta}(\mathrm{F}_j)$.

Hence, $0 \leq 1 + \mathrm{NIG}_{\delta}(F_i, F_j)/(\mathrm{NH}_{\delta}(F_i) + \mathrm{NH}_{\delta}(F_j)) \leq 2$.

By Definition 16, we have $0 \leq \mathrm{NIW}_{\delta}(F_i, F_j) \leq 2$. □

Theorem 18. *If feature F_i is redundant with feature F_j, then* $0 \leq NIW_{\delta}(F_i, F_j) \leq 1$.

Proof. According to Definition 14, if feature F_i is redundant with feature F_j, then $\mathrm{NIG}_{\delta}(F_i, F_j) \leq 0$. It is known that $\mathrm{NIG}_{\delta}(F_i, F_j) \geq -[\mathrm{NH}_{\delta}(F_i) + \mathrm{NH}_{\delta}(F_j)]$, and then $-1 \leq \mathrm{NIG}_{\delta}(F_i, F_j)/(\mathrm{NH}_{\delta}(F_i) + \mathrm{NH}_{\delta}(F_j)) \leq 0$. Hence, $0 \leq \mathrm{NIW}_{\delta}(F_i, F_j) \leq 1$. □

Theorem 19. *If feature F_i is interactive with feature F_j, then* $1 \leq NIW_{\delta}(F_i, F_j) \leq 2$.

Proof. According to Definition 15, if feature F_i is interactive with feature F_j, then $\mathrm{NIG}_{\delta}(F_i, F_j) \geq 0$. It is known that $\mathrm{NIG}_{\delta}(F_i, F_j) \leq \mathrm{NH}_{\delta}(F_i) + \mathrm{NH}_{\delta}(F_j)$, and then $0 \leq \mathrm{NIG}_{\delta}(F_i, F_j)/(\mathrm{NH}_{\delta}(F_i) + \mathrm{NH}_{\delta}(F_j)) \leq 1$. Hence, $1 \leq \mathrm{NIW}_{\delta}(F_i, F_j) \leq 2$. □

4.2. Proposed Feature Selection Algorithm. Previous feature selection algorithms seldom consider redundancy and interaction at the same time. This results in loss of some valuable features in the process of feature selection. To solve this problem, we first compute the neighborhood mutual information between a feature and the class and then adjust it through the manipulation of interaction weight factor which can reflect the information of whether a feature is redundant or interactive. The candidate features will be ranked with the adjusted relevance measure. The corresponding descriptive pseudocode is shown in Algorithm 1.

Features can be selected by different search strategies. For the sake of efficiency, we use the sequential forward search technique in this paper. A predefined threshold K is used to terminate the procedure, and δ is a neighborhood parameter. For a dataset D with original set $\mathbb{F} = \{F_1, F_2, \ldots, F_n\}$ and the class C, we rank the features in the descending order according to the adjusted relevance measure and then select the first K features, where K has been specified in advance.

NIWFS is a feature ranking algorithm. Firstly, we initialize parameters which consist of the selected feature subset and the weight for each feature and employ the neighborhood mutual information $\mathrm{NMI}_{\delta}(F_i; C)$ as a measure of relevance. Secondly, candidate features will be weighted through the neighborhood interaction weight factor $\mathrm{NIW}_{\delta}(F_i, F_j)$. And the original relevance $\mathrm{NMI}_{\delta}(F_i; C)$ will be redressed by multiplying the weight $w(F_i)$. Feature F_j with the largest $R_{\delta}(F_i; C)$ will be selected and removed from the feature set $\mathbb{F}$ to subset S. This process terminates until K features have been selected. According to Theorem 18, the weight of a redundant feature is in the range of $[0, 1]$, and the value of $R_{\delta}(F_i; C)$ will decrease by multiplying the weight $w(F_i)$. According to Theorem 19, the weight of an interactive feature is in the range of $[1, 2]$, and the value of $R_{\delta}(F_i; C)$ will increase by multiplying the weight $w(F_i)$. Therefore, the adjusted relevance measure can reflect the information of whether a feature is redundant or interactive.

To determine the threshold K, we may use a specific classifier to select the subset of features producing the highest accuracy. Alternatively, we may terminate the procedure until $|\mathrm{NMI}_{\delta}(\mathbb{F}; C) - \mathrm{NMI}_{\delta}(S; C)| \leq \varepsilon$ is satisfied, where ε is a very little positive number.

Now we analyze the complexity of the algorithm. Let us suppose that n is the number of candidate features in the given

```
Input: Dataset $D$ with original feature set $\mathbb{F} = \{F_1, F_2, \dots, F_n\}$ and the class $C$
       Number of selected feature $K$, neighborhood size $\delta$
Output: Selected feature subset $S$
$S \leftarrow \varnothing$;
$k \leftarrow 0$;
Initial the weight $w(F_i)$ to 1 for each feature;
For each $F_i \in \mathbb{F}$ do
    Calculate $\mathrm{NMI}_\delta(F_i; C)$ using Definition 8;
End
While $k < K$ do
    For each candidate feature $F_i \in \mathbb{F}$ do
        Calculate the adjusted relevance measure $R_\delta(F_i; C) = w(F_i) \times \mathrm{NMI}_\delta(F_i; C)$;
    End
    Select the feature $F_j$ with the largest $R_\delta(F_i; C)$;
    $S = S \cup \{F_j\}$;
    $\mathbb{F} = \mathbb{F} - \{F_j\}$;
    For each candidate feature $F_i \in \mathbb{F}$ do
        Calculate the interaction weight factor $\mathrm{NIW}_\delta(F_i, F_j)$;
        Update $w(F_i) = w(F_i) \times \mathrm{NIW}_\delta(F_i, F_j)$;
    End
    $k = k + 1$;
End
```

Algorithm 1: NIWFS: neighborhood interaction weight based feature selection algorithm.

dataset D. First of all, we need to calculate the neighborhood mutual information between n features and the class, and the time complexity is $O(n)$. We assume that K features have been selected and compute the adjusted relevance measure between the $n - K$ remaining features and the class. The computational complexity is $\sum_{i=1}^{K}(n - i + 1)$. Besides, the time complexity of calculating updated weight is $\sum_{i=2}^{K}(n - i + 1)$. Therefore, the total complexity of NIWFS is $O(nK)$, and it is the same as that of NMI. In the worst case, the total complexity is $O(n^2)$ when all features are selected. However, in most cases, $K \ll n$.

5. Experiments

In this section, we empirically evaluate the performance of our proposed algorithm and present the experimental results in comparison with the other three different types of feature subset selection algorithms applied to ten real world datasets, respectively.

5.1. Experiment Setup. To verify the effectiveness of our method, ten datasets are downloaded from UCI machine learning repository [43]. The description of datasets is presented in Table 2. Among 10 datasets, three are completely discrete, three are completely continuous, and the other four are heterogeneous. The sizes of datasets vary from 32 to 8124, the numbers of candidate features vary from 13 to 279, and the classes vary from 2 to 19.

All the continuous features are transformed to interval $[0, 1]$ in preprocessing, while the discrete features are coded with a sequence of integers. The 2-norm is used to compute distance (Euclidean distance). The neighborhood parameter δ is set as 0 for the datasets with categorical features. According to observations made by Hu et al. [31], the threshold δ should take value in [0.1, 0.2] for numerical features. In the following, we set the neighborhood parameter to 0.15 as suggested by Hu et al. [35]. As INTERACT cannot deal with numerical features directly, we employ the MDL discretization method to transform the numerical features into discrete one [44]. For datasets with missing values, we replace all missing values for nominal and numerical features with the modes and means from the training data [45].

Three representative feature selection algorithms are selected to be compared with NIWFS. To evaluate the performance of NIWFS in terms of handling feature interaction, an algorithm INTERACT [16], which is specifically proposed to address the feature interaction, is selected as one benchmark algorithm. Moreover, we also compare NIWFS with NRS and NMI which can handle mixed datasets directly. NRS evaluates the features with a function called dependency, which is the ratio of consistent samples over the whole learning samples; NMI employs the neighborhood mutual information to select relevant features based on the criterion of maximal dependency.

We use specific classifier to select the top K features producing the highest accuracy. Two representative classification algorithms based on different hypotheses are employed to test the performance of selected features. They are tree-based C4.5 [46] and instance-based IB1 [47], respectively. The whole classification process is conducted in WEKA release 3.6.9 with default parameter settings.

5.2. Experimental Results and Analysis. The classification accuracy is obtained by 10-fold cross-validation. The results

Table 2: Experimental datasets description.

Number	Datasets	Instances	Total features	Numerical features	Categorical features	Classes
1	Arrhythmia	452	279	206	73	13
2	Autos	205	25	15	10	6
3	Credit-g	1000	20	6	14	2
4	Heart-c	303	13	6	7	5
5	Lung-cancer	32	56	0	56	2
6	Movement_libras	360	90	90	0	15
7	Mushroom	8124	22	0	22	2
8	Sonar	208	60	60	0	2
9	Soybean	683	35	0	35	19
10	Synthetic_control	600	60	60	0	6

Table 3: Number and accuracy (%) of features selected with different algorithms (C4.5).

Number	Unselect		INTERACT		NRS		NMI		NIWFS	
	Acc.	n	Acc.	n	Acc.	n	Acc.	n	Acc.	n
1	63.9	279	62.6^{*}	22	66.6^{*}	28	65.0^{*}	15	**69.2**	19
2	78.0	25	78.5	8	79.0	14	80.5	12	**81.0**	7
3	70.8	20	74.6^{*}	13	74.5^{*}	5	72.9^{*}	13	**75.4**	5
4	76.6	13	79.2^{*}	10	80.2^{*}	4	81.2	5	**81.8**	3
5	50.0	56	65.6^{*}	6	**75.0**	3	**75.0**	2	**75.0**	2
6	69.7	90	64.4^{*}	18	67.5	28	65.6^{*}	28	**67.8**	19
7	100.0	22	**100**	5	**100.0**	5	**100.0**	4	**100.0**	6
8	71.2	60	76.9	12	73.1^{*}	18	76.4^{*}	20	**77.9**	16
9	92.4	35	84.3^{*}	12	91.1	19	90.6	20	**91.5**	17
10	91.5	60	79.5^{*}	16	91.3	11	92.0	17	**92.8**	12
Avg.	76.4	66	76.5	12.2	79.8	13.5	79.9	13.6	**81.2**	10.6
WTL			0/3/7		0/6/4		0/6/4			

in Tables 3 and 4 show that the classification accuracies and the number of the selected features obtained by the original features (Unselect), INTERACT, NRS, NMI, and NIWFS with different classifiers. The bold value means that it is the largest one among these four feature selection algorithms. The row Avg. shows the average of accuracies and the number of selected features with different learning algorithms. In addition, a paired two-tailed t-test between accuracies of NIWFS and other selectors has been performed. Moreover, the number of the datasets which have higher (or equal or lower) accuracy with respect to NIWFS is represented by the WTL (win/tie/loss). The symbols "v" and "$*$," respectively, identify statistically significant (at 0.05 level) wins or losses over our proposed method.

As we can see in Tables 3 and 4, all of the feature selection algorithms can remove a large number of candidate features effectively. Our proposed NIWFS algorithm obtains the best average accuracies for all the classification algorithms. For instance, with respect to IB1 learning algorithm, the average accuracy is 84.7% for NIWFS, while accuracies of INTERACT, NRS, and NMI are 78.7%, 81.4%, and 81.0%, respectively. The average classification accuracy reduced by 7.6%, 4.1%, and 4.6%, respectively. By comparing the accuracies of the four feature selection algorithms, we can find that NIWFS exhibits the highest classification accuracy.

The results obtained by NIWFS method are better than or at least equal to those obtained by the INTERACT, NRS, and NMI methods according to the view of win/tie/loss. For example, the numbers of cases for which NIWFS achieves significantly higher classification accuracy over INTERACT, NRS, and NMI are seven, seven, and eight out of ten cases in the IB1 classifier, respectively.

The average numbers of selected features achieved by NIWFS are 10.6 and 12.6, respectively. Notice that the average number of selected feature obtained by NIWFS is 10.6 in the C4.5 classifier which is the least among these methods. In general, our proposed algorithm achieves better results as compared with the other three feature selection algorithms.

From the experimental results, we also find that INTERACT has the lowest average classification accuracy among the four feature selection algorithms. There are two main reasons for this. Firstly, INTERACT can only deal with nominal features. Therefore, some valuable information may be lost in the process of discretization. Secondly, INTERACT does not use wrapper. The reason why NIWFS wins over NMI and NRS is that NIWFS considers not only the relevance between a single feature and the class, but also the redundancy and interaction with other features which are expressed by the interaction weight factor. Therefore, NIWFS performs better when there is feature interaction in the dataset.

Table 4: Number and accuracy (%) of features selected with different algorithms (IB1).

Number	Unselect		INTERACT		NRS		NMI		NIWFS	
	Acc.	*n*	Acc.	*n*	Acc.	*n*	Acc.	*n*	Acc.	*n*
1	53.8	279	55.5*	22	65.3	25	61.3*	19	**66.4**	23
2	72.2	25	70.2*	8	78.5*	8	75.6*	15	**83.9**	11
3	70.5	20	69.5*	13	71.7	11	70.3*	14	**71.9**	6
4	75.9	13	76.6*	10	77.9*	4	78.9	4	**80.2**	5
5	37.5	56	81.3	6	75.0*	4	78.1*	6	**84.4**	9
6	85.8	90	83.9*	18	84.7*	27	83.6*	29	**86.7**	17
7	100.0	22	**100.0**	5	**100.0**	5	**100.0**	4	**100.0**	6
8	86.5	60	**87.5**	12	82.2*	20	83.2*	17	**87.5**	17
9	91.7	35	79.2*	12	86.7*	20	87.1*	20	**89.6**	14
10	96.5	60	83.5*	16	91.7*	8	91.5*	7	**96.3**	18
Avg.	77.0	66	78.7	12.2	81.4	13.2	81.0	13.5	**84.7**	12.6
WTL			0/3/7		0/3/7		0/2/8			

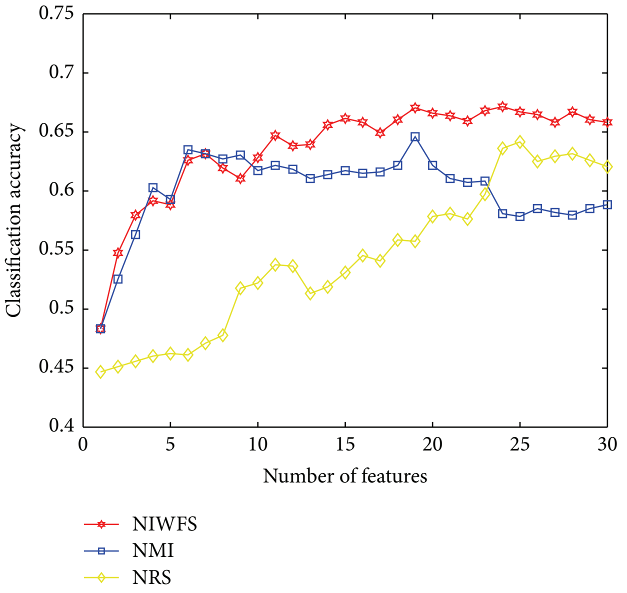


Figure 1: Average classification accuracy versus different number of selected features on arrhythmia dataset.

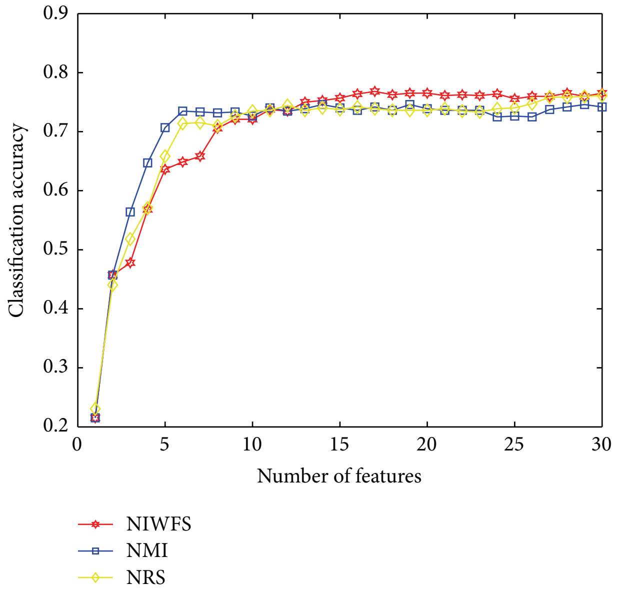


Figure 2: Average classification accuracy versus different number of selected features on movement_libras dataset.

To further compare the effectiveness of NIWFS with NMI and NRS which can deal with mixed data directly, we add features for learning one by one in the order that the features are selected. In the experiments, three representative datasets are chosen: arrhythmia, movement_libras, and synthetic_control. To reduce the bias of a feature assessment based on a specific classification, we calculate the average classification accuracies of classifiers for NIWFS, NMI, and NRS. The comparison results are shown as in Figures 1, 2, and 3. The number k in x-axis refers to the first k features with the selected order by different methods. The y-axis represents the average classification accuracies of the first k features.

The results in Figures 1–3 show that the best average accuracy of classifier with NIWFS is higher than NMI and NRS. On the arrhythmia dataset, the plots of NIWFS are much higher than NRS and higher than NMI in the range of 10–30 features. NIWFS achieves 67.15% classification accuracy with 24 features, which is higher than NMI by 3.95% and higher than NRS by 4.68%. With the movement_libras dataset, the plots of NIWFS are higher than NMI and NRS in the range of 13–30 features. NIWFS produces its best accuracy (76.81% with 17 features), which is about 3% higher than NMI and about 1% higher than NRS. For the synthetic control dataset, the plots of NIWFS are much higher than NMI and NRS in the range of 9–20 features. The highest accuracy of synthetic_control achieved by NIWFS is 94.33% with 16 selected features while the highest accuracy is 91.92% with 7 selected features for NMI and 91.17% for NRS with 8 selected

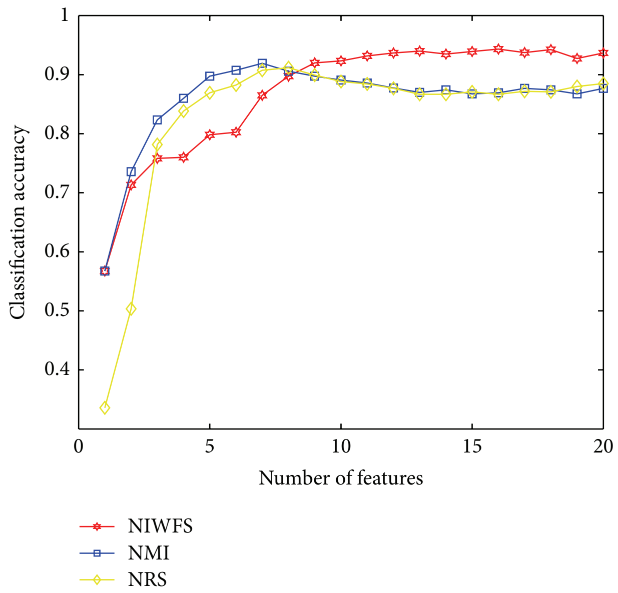


Figure 3: Average classification accuracy versus different number of selected features on synthetic_control dataset.

features. This demonstrates that having too few features is not necessarily a good feature selection result. Some interactive features may be lost in the process of removing redundancy. We also find that plots with the first few features are lower than NMI and NRS in some cases. The main reason is that NIWFS does not select the first few features having the maximal relevance with the class due to the weight reducing by redundancy analysis.

6. Conclusions and Future Work

The main goal of feature selection is to find a feature subset that is small in size but high in prediction accuracy. Feature interaction exists in many applications. It is a challenging task to find interactive feature. In this paper, we present an interactive feature searching algorithm which is based on some neighborhood information measures. First, the new definitions of redundant and interactive feature have been defined in the framework of neighborhood rough sets. Then we propose the neighborhood interaction weight factor which can reflect the information of whether a feature is redundant or interactive. Based on the neighborhood interaction weight factor, we present our feature selection method. This method is compared with three other feature selection methods in terms of the number of selected features and accuracies of two classifiers such as C4.5 and IB1 on ten public real world datasets. The experimental results show that NIWFS can not only deal with mixed datasets directly, but also reduce a large number of features with the best average classification accuracies.

However, it is time-consuming for our method. The main reason is that the computation of the neighborhood mutual information involves the calculation of distance. For the future work, we plan to improve the efficiency of NIWFS.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Natural Science Foundation of China (70971137).

References

[1] I. Iguyon and A. Elisseeff, “An introduction to variable and feature selection,” *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[2] M. Dash and H. Liu, “Consistency-based search in feature selection,” *Artificial Intelligence*, vol. 151, no. 1-2, pp. 155–176, 2003.

[3] H. Liu, J. Sun, L. Liu, and H. Zhang, “Feature selection with dynamic mutual information,” *Pattern Recognition*, vol. 42, no. 7, pp. 1330–1339, 2009.

[4] H. Yang and J. Moody, “Feature selection based on joint mutual information,” in *Proceedings of the International ICSC Symposium on Advances in Intelligent Data Analysis*, pp. 22–25, 1999.

[5] F. Fleuret, “Fast binary feature selection with conditional mutual information,” *The Journal of Machine Learning Research*, vol. 5, pp. 1531–1555, 2004.

[6] R. Cai, Z. Hao, X. Yang, and W. Wen, “An efficient gene selection algorithm based on mutual information,” *Neurocomputing*, vol. 72, no. 4–6, pp. 991–999, 2009.

[7] G. Wang and F. H. Lochovsky, “Feature selection with conditional mutual information maximin in text categorization,” in *Proceedings of the 13th International Conference on Information and Knowledge Management (CIKM ’04)*, pp. 342–349, ACM, 2004.

[8] D. Levi and S. Ullman, “Learning to classify by ongoing feature selection,” *Image and Vision Computing*, vol. 28, no. 4, pp. 715–723, 2010.

[9] M. Robnik-Šikonja and I. Kononenko, “Theoretical and empirical analysis of ReliefF and RReliefF,” *Machine Learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[10] I. Kononenko, “Estimating attributes: analysis and extensions of RELIEF,” in *Machine Learning: ECML-94*, pp. 171–182, Springer, Berlin, Germany, 1994.

[11] M. A. Hall, *Correlation-based feature selection for machine learning [Ph.D. thesis]*, The University of Waikato, 1999.

[12] R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.

[13] L. Yu and H. Liu, “Efficient feature selection via analysis of relevance and redundancy,” *Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2003/04.

[14] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[15] A. Jakulin and I. Bratko, “Analyzing attribute dependencies,” in *Proceedings of the 7th European Conference on Principles*

and Practice of Knowledge Discovery in Databases, pp. 229–240, September 2003.

[16] Z. Zhao and H. Liu, "Searching for interacting features in subset selection," *Intelligent Data Analysis*, vol. 13, no. 2, pp. 207–228, 2009.

[17] G. Wang, Q. Song, B. Xu, and Y. Zhou, "Selecting feature subset for high dimensional data via the propositional FOIL rules," *Pattern Recognition*, vol. 46, no. 1, pp. 199–214, 2013.

[18] Q. Hu, J. Liu, and D. Yu, "Mixed feature selection based on granulation and approximation," *Knowledge-Based Systems*, vol. 21, no. 4, pp. 294–304, 2008.

[19] R. Jenson and Q. Shen, "Fuzzy-rough sets for descriptive dimensionality reductions," in *Proceedings of the IEEE International Conference on Fuzzy Systems*, pp. 29–34, 2002.

[20] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991.

[21] Z. Pawlak, "Vagueness and uncertainty: a rough set perspective," *Computational Intelligence*, vol. 11, no. 2, pp. 227–232, 1995.

[22] Z. Pawlak, "Rough set approach to knowledge-based decision support," *European Journal of Operational Research*, vol. 99, no. 1, pp. 48–57, 1997.

[23] Z. Pawlak, "Rough set theory and its applications to data analysis," *Cybernetics and Systems*, vol. 29, no. 7, pp. 661–688, 1998.

[24] T. Herawan, M. M. Deris, and J. H. Abawajy, "A rough set approach for selecting clustering attribute," *Knowledge-Based Systems*, vol. 23, no. 3, pp. 220–231, 2010.

[25] N. M. Parthaláin and Q. Shen, "Exploring the boundary region of tolerance rough sets for feature selection," *Pattern Recognition*, vol. 42, no. 5, pp. 655–667, 2009.

[26] J.-S. Mi, W.-Z. Wu, and W.-X. Zhang, "Approaches to knowledge reduction based on variable precision rough set model," *Information Sciences*, vol. 159, no. 3-4, pp. 255–272, 2004.

[27] R. W. Swiniarski and A. Skowron, "Rough set methods in feature selection and recognition," *Pattern Recognition Letters*, vol. 24, no. 6, pp. 833–849, 2003.

[28] K. Thangavel and A. Pethalakshmi, "Dimensionality reduction based on rough set theory: a review," *Applied Soft Computing Journal*, vol. 9, no. 1, pp. 1–12, 2009.

[29] C. Yang, W. Zhang, J. Zou, S. Hu, and J. Qiu, "Feature selection in decision systems: a mean-variance approach," *Mathematical Problems in Engineering*, vol. 2013, Article ID 268063, 8 pages, 2013.

[30] Q. Hu, Z. Xie, and D. Yu, "Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation," *Pattern Recognition*, vol. 40, no. 12, pp. 3509–3521, 2007.

[31] Q. Hu, D. Yu, J. Liu, and C. Wu, "Neighborhood rough set based heterogeneous feature subset selection," *Information Sciences*, vol. 178, no. 18, pp. 3577–3594, 2008.

[32] Q. Hu, D. Yu, and Z. Xie, "Information-preserving hybrid data reduction based on fuzzy-rough techniques," *Pattern Recognition Letters*, vol. 27, no. 5, pp. 414–423, 2006.

[33] R. Jensen and Q. Shen, "Fuzzy-rough sets assisted attribute selection," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 1, pp. 73–89, 2007.

[34] R. Jensen and Q. Shen, "New approaches to fuzzy-rough feature selection," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 4, pp. 824–838, 2009.

[35] Q. Hu, L. Zhang, D. Zhang, W. Pan, S. An, and W. Pedrycz, "Measuring relevance between discrete and continuous features based on neighborhood mutual information," *Expert Systems with Applications*, vol. 38, no. 9, pp. 10737–10750, 2011.

[36] Q. Hu, D. Yu, and Z. Xie, "Neighborhood classifiers," *Expert Systems with Applications*, vol. 34, no. 2, pp. 866–876, 2008.

[37] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[38] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, NY, USA, 1991.

[39] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Proceedings of the 11th International Conference on Machine Learning*, pp. 121–129, 1994.

[40] A. Jakulin and I. Bratko, "Testing the significance of attribute interactions," in *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*, pp. 409–416, ACM, July 2004.

[41] A. Jakulin, *Attribute interactions in machine learning [M.S. thesis]*, Computer and Information Science, University of Ljubljana, 2003.

[42] W. J. McGill, "Multivariate information transmission," *Psychometrika*, vol. 19, no. 2, pp. 97–116, 1954.

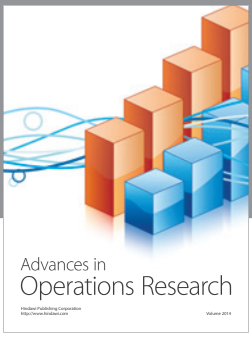
[43] A. Asuncion and D. Newman, "UCI machine learning repository," 2007.

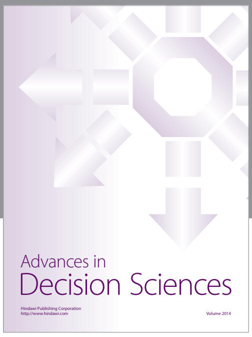
[44] U. Fayyad and K. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proceedings of 13th International Joint Conference on Artificial Intelligence*, pp. 1022–1027, 1993.

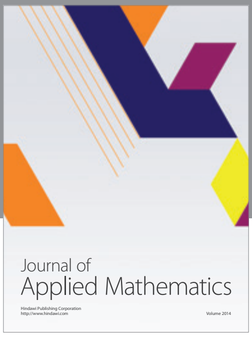
[45] J. W. Grzymala-Busse and M. Hu, "A comparison of several approaches to missing attribute values in data mining," in *Rough Sets and Current Trends in Computing*, pp. 378–385, Springer, Berlin, Germany, 2001.

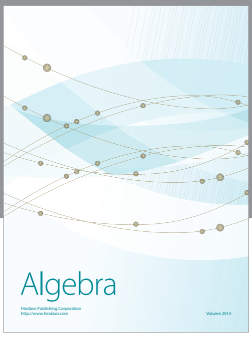
[46] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

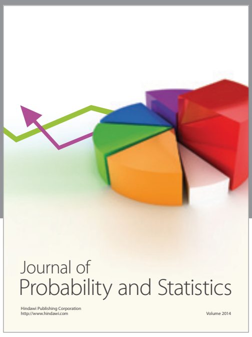
[47] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, no. 1, pp. 37–66, 1991.


Advances in
Operations Research

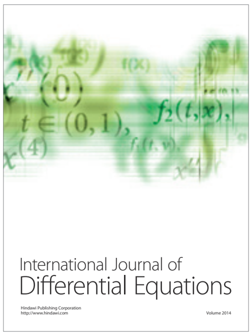

Advances in
Decision Sciences


Journal of
Applied Mathematics

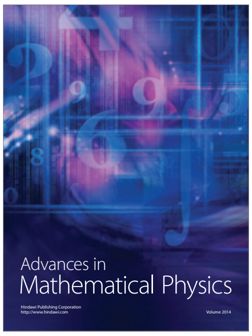

Algebra

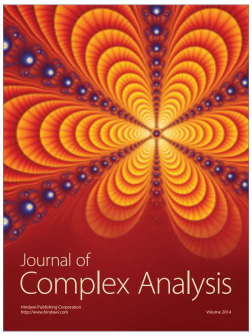

Journal of
Probability and Statistics

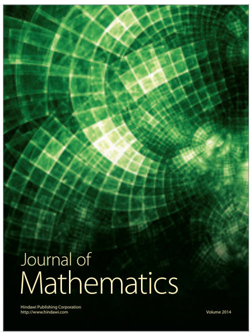
The Scientific
World Journal

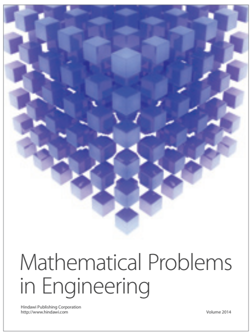

International Journal of
Differential Equations

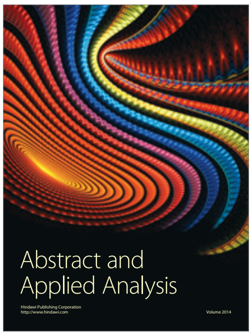
International Journal of
Combinatorics

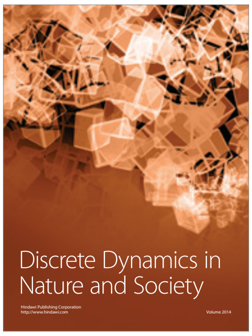

Advances in
Mathematical Physics

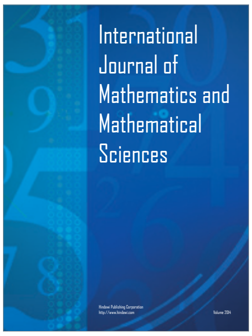

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

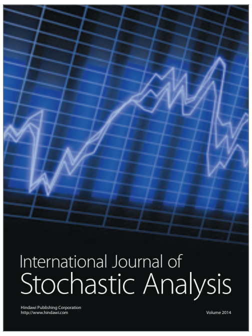

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization