Feature Selection for Monotonic Classification

Article *in* IEEE Transactions on Fuzzy Systems · July 2012
DOI: 10.1109/TFUZZ.2011.2167235 · Source: DBLP

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

Feature Selection for Monotonic Classification

Qinghua Hu, *Member, IEEE,* Weiwei Pan, Lei Zhang, *Member, IEEE,* David Zhang, *Fellow, IEEE,* Yanping Song, Maozu Guo, and Daren Yu

***Abstract*—Monotonic classification is a kind of special tasks in machine learning and pattern recognition. Monotonicity constraints between features and decision should be taken into account in these tasks. However, most of existing techniques are not able to discover and represent the ordinal structures in monotonic datasets. Thus they are inapplicable to monotonic classification. Feature selection has been proven effective in improving classification performance and avoiding overfitting. To the best of our knowledge, no technique has been specially designed for selecting features in monotonic classification till now. In this work, we introduce a function, called rank mutual information, to evaluate monotonic consistency between features and decision in monotonic tasks. This function combines the advantages of dominance rough sets in reflecting ordinal structures and mutual information in terms of robustness. Then rank mutual information is integrated with the search strategy of min-Redundancy and Max-Relevance for computing optimal subsets of features. A collection of numerical experiments are given to show the effectiveness of the proposed technique.**

***Index Terms*—monotonic classification, feature selection, fuzzy ordinal set, rank mutual information**

I. Introduction

CLASSIFICATION tasks can be divided into two groups: nominal classification and ordinal classification. As to nominal classification [52], [53], [56], there is no ordinal structure among different decision values. For example, we recognize different diseases according to the symptoms of patients. However, as to ordinal classification (also called ordinal regression) [1], [3], [4], [55], we should consider the ordinal relationship between different class labels, such as the severity levels of a disease {slight, medium, severe}. Furthermore, monotonic classification is a class of special ordinal classification tasks, where the decision values are ordinal and discrete, and there are monotonicity constraints between features and decision classes that $x \leq x' \Rightarrow f(x) \leq f(x')$ [1]. Monotonic classification is a kind of common tasks in medical analysis, social and behavioral sciences [2]. Such problems have attracted increasing attention from the domains of machine learning and intelligence data analysis [3], [4], [5].

Q. H. Hu, W. W. Pan, Y. P. Song, M. Z. Guo and Daren Yu are with Harbin Institute of Technology, Harbin, China. E-mail: huqinghua@ hit.edu.cn.

L. Zhang and D. Zhang are with The Hong Kong Polytechnic University, Kowloon, Hong Kong, China. E-mail: cslzhang@comp.polyu.edu.hk

This work is supported by National Natural Science Foundation of China under Grant 60703013 and 10978011, Key Program of National Natural Science Foundation of China under Grant 60932008, National Science Fund for Distinguished Young Scholars under Grant 50925625 and China Postdoctoral Science Foundation. Dr. Hu is supported by The Hong Kong Polytechnic University (G-YX3B).

The previous work on monotonic classification can be roughly divided into two groups. One attempts to construct a theoretic framework for monotonic classification, including dominance rough set model [6], [7], [8], [9], [10] and ordinal entropy model [11]. While the other is dedicated to developing algorithms for learning decision models from samples [12], [13], [14], [15]. In 1999, Greco, et al. first introduced dominance relations into rough sets and proposed the model of dominance rough sets. This model built a formal framework for studying monotonic classification. After that, this model was extensively discussed and generalized. on the other hand, Ben-David extended the classical decision tree algorithm to monotonic classification in 1995. Since then, a collection of decision tree algorithms have been developed for this problem [16], [17], [18], [19], [20]. In addition, Ben-David et al. also extended the nearest neighbor classifier to monotonic tasks and designed an ordinal learning model (OLM) [22]. In 2003, Cao-Van introduced ordinal stochastic dominance learner (OSDL) based on associated cumulative distribution. In 2008, Lievens, Baets and Cao-Van presented a probabilistic framework which served as the base of instance-based algorithms for solving the supervised ranking problems [23]. Also in 2008, Duivesteijn and Feelders proposed a modified nearest neighbor algorithm for the construction of monotone classifiers from data by monotonizing training data. The relabeled data was subsequently used as the training set by a modified nearest neighbor algorithm [14]. Recently, support vector machines and other kernel machines are also adapted to such tasks [24], [25]. Based on the above survey, we can see that monotonic classification is becoming a hot topic in machine learning.

As we know, feature selection plays an important role in improving classification performance and speeding up training [26], [27]. A great number of feature selection algorithms have been designed for classification learning till now. The main differences between these techniques lie in the metrics used to evaluate the quality of candidate features and search strategies for finding optimal solutions in terms of the used metric. Mutual information [28], [29], [30], [31], dependency [32], [33], [34], [35], [36], consistency [37], [38], [54], distance [39], and classification margin [40], [41], [42] were introduced or developed as metrics of feature quality in feature selection. In addition, after defining the optimization objectives, a search strategy should be designed to find the optimal solution. Greedy search, heuristic search, branch and bound, genetic optimization and other intelligent search algorithms are used in feature selection [43], [44], [45].

Although a lot of algorithms were developed for feature selection, little effort has been devoted to design feature selection algorithms for monotonic classification yet. As different

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

consistency assumptions are taken for monotonic classification and nominal classification, the feature evaluation functions developed for nominal classification cannot be directly applied to monotonic classification because the metrics in nominal classification do not consider the monotonicity constraints. As a result, a feature producing a large value of feature quality may not be useful for enhancing the monotonicity of monotone tasks. Thus new evaluation functions should be developed for this kind of special tasks. Kamishima and Akaho in 2006 [46] and Baccianella et al. in 2010 [47] designed a feature selection procedure for ordinal classification, respectively. However, the feature evaluation functions used in these algorithms do not reflect the monotonicity between features and decision. So they are not applicable to monotonic classification. In 2006, Lee, Yeung and Tsang improved the dependency function defined in dominance rough sets and used it to attribute reduction for monotonic classification. In addition, Xu, Zhang, Zhong et al. gave another framework of attribute reduction based on evidence theory [48]. Although dependency defined in dominance rough sets can reflect the ordinal structures in monotonic data, dominance rough sets are very sensitive to noisy information. The evaluation function may vary much if there are several inconsistent samples in the datasets [49]. We should design a robust metric of feature quality, which can also discover ordinal structures of monotone tasks.

In 2010, Hu, Guo and Yu introduced two new attribute metrics, called rank mutual information and fuzzy rank mutual information, to compute the monotonic consistency between two random variables [11]. However, they did not discuss the issue of feature selection for monotonic classification. Also, no experimental analysis was described to show the effectiveness of the proposed measure. As we know, mutual information in Shannon's information theory is widely used in feature evaluation for nominal classification tasks and its effectiveness has been verified in applications [26], [27], [28], [29], [30], [50]. Naturally, we desire rank mutual information and fuzzy rank mutual information will also be powerful in evaluating and selecting monotonic features. So in this work, we first discuss the properties of rank entropy and rank mutual information in evaluating features, and then we design feature selection algorithms based on these metrics and conduct experiments to test them. We integrate rank mutual information with the search strategy of min-Redundancy and Max-Relevance (mRMR). Thus an effective algorithm for monotonic feature selection is constructed. Some numerical experiments are presented to show the effectiveness of the proposed technique.

The rest of the paper is organized as follows. First we present the preliminaries on monotonic classification and dominance rough sets in Section 2; then we show the definitions of rank mutual information and fuzzy rank mutual information, and discuss their properties in Section 3. Section 4 gives the feature selection algorithms for monotonic classification. Numerical experiments are presented in Section 5. Finally conclusions and future work are given in Section 6.

II. Preliminaries on monotonic classification

The following definitions can be found in [6], [7].

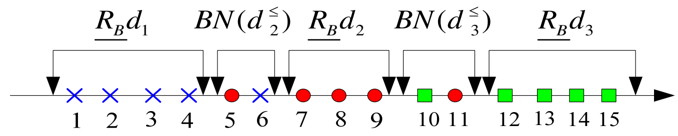


Fig. 1. Toy example of dominance rough sets

Definition 1: Let $< U, A, D >$ be a set of classification dataset, where $U = \{x_i\}_{i=1}^n$ is the set of samples, $A = \{a_j\}_{j=1}^m$ is the set of attributes and D is the decision of the samples. The value domain of D is $\{d_1, d_2, \cdots, d_K\}$. If D is nominal, we say $< U, A, D >$ is a nominal classification task. If there are ordinal structures between the values of decision, $d_1 < d_2 < \cdots < d_K$, we say $< U, A, D >$ is an ordinal classification task. Let $v(x, A)$ denote the value vector of sample x on A, f be the decision function. If $\forall x \in U, v(x, A) \leq v(x', A)$, we have $f(x) \leq f(x')$, then we say $< U, A, D >$ is a monotonic classification task.

In this work, we focus on monotonic classification tasks.

Definition 2: Given a monotonic classification task $< U, A, D >$, $B \subseteq A$, we associate ordinal relations with the attributes and decision as

1) $R_B^{\leq} = \{(x_i, x_j) | v(x_i, a_l) \leq v(x_j, a_l), \forall a_l \in B\}$;
2) $R_D^{\leq} = \{(x_i, x_j) \in U \times U | v(x_i, D) \leq v(x_j, D)\}$.

Definition 3: Given $< U, A, D >$, $B \subseteq A$. $\forall x_i \in U$, we define the following subsets of samples:

1) $[x_i]_B^{\leq} = \{x_j \in U | (x_i, x_j) \in R_B^{\leq}\}$,
2) $[x_i]_D^{\leq} = \{x_j \in U | (x_i, x_j) \in R_D^{\leq}\}$,

called B-dominating set and D-dominating set of x_i, respectively.

Given $< U, A, D >$, $B \subseteq A$, $x_i, x_j \in U$, the following conclusions hold: 1) $R_A^{\leq} \subseteq R_B^{\leq}$; 2) $[x_i]_A^{\leq} \subseteq [x_i]_B^{\leq}$; 3) If $x_j \in [x_i]_B^{\leq}$, then $[x_j]_B^{\leq} \subseteq [x_i]_B^{\leq}$ and $[x_i]_B^{\leq} = \cup\{[x_j]_B^{\leq} | x_j \in [x_i]_B^{\leq}\}$.

Definition 4: Given $< U, A, D >$, $B \subseteq A$, $X \subseteq U$. The lower and upper approximations of X in terms of B are defined as

1) $\underline{R_B^{\leq}}X = \{x \in U | [x]_B^{\leq} \subseteq X\}$;
2) $\overline{R_B^{\leq}}X = \{x \in U | [x]_B^{\geq} \cap X \neq \emptyset\}$.

It is easy to obtain the following conclusions: 1) $\underline{R_B^{\leq}}X \subseteq X \subseteq \overline{R_B^{\leq}}X$; 2) $\underline{R_B^{\leq}}U \subseteq U$, $\overline{R_B^{\leq}}\emptyset \subseteq \emptyset$; 3) $\underline{R_B^{\leq}} \sim X = \sim \overline{R_B^{\leq}}X$, $\overline{R_B^{\leq}} \sim X = \sim \underline{R_B^{\leq}}X$; 4) if $X \subseteq Y \subseteq U$, $\underline{R_B^{\leq}}X \subseteq \underline{R_B^{\leq}}Y$, $\overline{R_B^{\leq}}X \subseteq \overline{R_B^{\leq}}Y$.

Definition 5: Given $< U, A, D >$, $B \subseteq A$. d_i is the i'th class, the boundary of d_i are defined as

$$BN(d_i^{\leq}) = \overline{R_B^{\leq}}d_i^{\leq} - \underline{R_B^{\leq}}d_i^{\leq}. \tag{1}$$

Definition 6: Given $< U, A, D >$, $B \subseteq A$, and $\{d_1, d_2, \cdots, d_K\}$ is the value domain of D. The boundary of classification D is defined as

$$BN(D^{\leq}) = \bigcup_{i=1}^{K} BN(d_i^{\leq}). \tag{2}$$

Similarly, we can also define $BN(d_i^{\geq})$ and $BN(D^{\geq})$. It is easy to derive that $BN(d_i^{\leq}) = BN(d_{i+1}^{\geq})$ and $BN(D^{\geq}) = BN(D^{\leq})$. Moreover, we define $\underline{R_B d_i} = \underline{R_B^{\leq}}d_i^{\leq} \cap \underline{R_B^{\geq}}d_i^{\geq}$.

Example 1: A set of objects are divided into three levels according to the attribute B, presented in Fig. 1, where $\times, \bullet, \square$ stand for samples coming from classes 1, 2, and 3, respectively.

According to the above definitions, we obtain that $\underline{R_B}d_1 = \{x_1, x_2, x_3, x_4\}$, $BN(d_2^{\leq}) = \{x_5, x_6\}$, $\underline{R_B}d_2 = \{x_7, x_8, x_9\}$, $BN(d_2^{\leq}) = \{x_{10}, x_{11}\}$, $\underline{R_B}d_3 = \{x_{12}, x_{13}, x_{14}, x_{15}\}$.

The samples in the boundary set are the source of difficulty of classification. They form the inconsistency of classification. We give a metric to evaluate the monotonic consistency as

$$\gamma_B(D) = \frac{|U - \bigcup_{i=1}^{K} BND_B(d_i)|}{|U|}, \tag{3}$$

where $|X|$ is the number of the elements in X. We call this metric monotonic dependency of D on B. If $\gamma_B(D) = 1$, we say D is completely dependent on B. The dataset is monotonically consistent in this case. All the samples with better feature values also obtain better decision labels. However, most classification tasks are not consistent in real-world applications.

As to the task in Example 1, we have $\gamma_B(D) = |U - \{x_5, x_6\} \cup \{x_{10}, x_{11}\}|/|U| = 11/15$. Monotonic dependency characterizes the relevance between attributes and classification. However, this metric is sensitive to noisy samples. For example, we just change the decision of Sample 15 as class 1, then $BN(D^{\leq}) = \{x_5, \cdots, x_{14}\}$ and $\gamma_B(D) = |U - \{x_5, x_6\} \cup \{x_{10}, x_{11}\}|/|U| = 5/15$. As we know the decisions are usually given by different persons in different contexts; there are many inconsistent decisions in data, so a robust metric is desirable in this case.

Due to inconsistency, a sample with higher values of features does not necessarily obtain a better decision. However, we know that a sample with larger values of features should produce a better decision with a large probability [4], [5]. For applicability, stochastic monotonicity should be considered to describe monotonic classification tasks.

III. Monotonic consistency metric

There are several kinds of uncertainty in monotonic classification, such as randomicity, fuzziness and inconstancy. The metric for evaluating quality of features should consider these problems. First, we introduce some definitions on rank entropy and rank mutual information [11], which reflects the stochastic monotonicity between features and decision.

Definition 7: Let $U = \{x_1, x_2, \cdots, x_n\}$ and $x_i \in U$ and $R = \{r_{ij}\}_{n \times n}$ be an ordinal relation over U. The fuzzy ordinal set of x_i is formulated as $[x_i]_R^{\leq} = r_{i1}/x_1 + r_{i2}/x_2 + \cdots + r_{in}/x_n$, where r_{ij} is the degree of x_i worse than x_j. We have

$$\begin{cases} x_i > x_j, \ r_{ij} \in [0, 0.5) \\ x_i = x_j, \ r_{ij} = 0.5 \\ x_i < x_j, \ r_{ij} \in (0.5, 1] \end{cases}. \tag{4}$$

The fuzzy dominated set of x_i is a fuzzy set which dominates x_i, The membership r_{ij} reflects the magnitude of x_i worse than x_j.

If we define a cut operator on the above fuzzy ordinal set as $r_{ij} = 0$ if $r_{ij} < 0.5$; otherwise, $r_{ij} = 1$, then the fuzzy ordinal set becomes a crisp ordinal set, as shown in Fig. 2.

Definition 8: Let U be a set of objects and $R = \{r_{ij}\}_{n \times n}$ be an ordinal relation over U induced by $B \subseteq A$. $[x_i]_R^{\leq}$ is the fuzzy ordinal set associated with x_i. The fuzzy rank entropy of the system $< U, R >$ is defined as

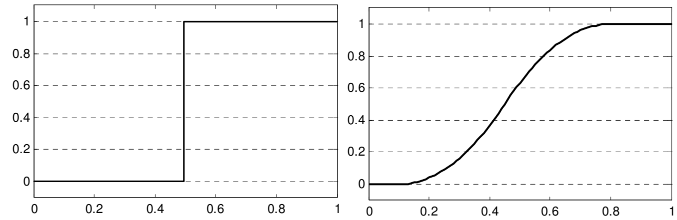


Fig. 2. Membership functions of ordinal sets

$$RH_R(U) = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{|[x_i]_R^{\leq}|}{n}, \tag{5}$$

where $|[x_i]_R^{\leq}| = \sum_j r_{ij}$ is the fuzzy cardinality of fuzzy set $[x_i]_R^{\leq}$.

$RH_R(U)$ is also written as $RH_B(U)$. As we know $0 \leq |[x_i]_R^{\leq}| \leq n$, so $0 \leq RH_R(U)$. In addition, assume that R_1 and R_2 are two fuzzy ordinal relations on U. If $R_1 \subseteq R_2$, we have $RH_{R_1}(U) \geq RH_{R_2}(U)$.

Definition 9: Given U, R and S are two fuzzy ordinal relations on U induced by attributes B_1 and B_2. $T = R \cap S$ is the relation induced by $B_1 \cup B_2$. That is to say $[x_i]_T^{\leq} = [x_i]_R^{\leq} \cap [x_i]_S^{\leq} = \min(r_{i1}, s_{i1})/x_1 + \min(r_{i2}, s_{i2})/x_2 + \cdots + \min(r_{in}, s_{in})/x_n$. The fuzzy rank joint entropy of R and S is defined as

$$RH_{R \cap S}(U) = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{|[x_i]_R^{\leq} \cap [x_i]_S^{\leq}|}{n}. \tag{6}$$

It is easy to show that $RH_{R \cap S}(U) \geq 0$, $RH_{R \cap S}(U) \geq RH_R(U)$ and $RH_{R \cap S}(U) \geq RH_S(U)$. Moreover, if $R \subseteq S$, we have $RH_{R \cap S}(U) = RH_R(U)$. This analysis shows the joint entropy of two subsets of features is no smaller than the entropy of any of them. We can derive the following property. If $B \subseteq C$, we have $RH_B(U) \leq RH_C(U)$.

Definition 10: Given U, R and S are two fuzzy ordinal relations induced by attributes B_1 and B_2. Known B_2, the fuzzy rank conditional entropy of B_1 is defined as

$$RH_{R|S}(U) = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{|[x_i]_R^{\leq} \cap [x_i]_S^{\leq}|}{|[x_i]_S^{\leq}|}. \tag{7}$$

As $|[x_i]_S^{\leq}| \geq |[x_i]_R^{\leq} \cap [x_i]_S^{\leq}|$, $|[x_i]_R^{\leq} \cap [x_i]_S^{\leq}|/|[x_i]_S^{\leq}| \leq 1$, then we can derive that $RH_{R|S}(U) \geq 0$. In addition, $|[x_i]_R^{\leq} \cap [x_i]_S^{\leq}|/|[x_i]_S^{\leq}| \geq |[x_i]_R^{\leq} \cap [x_i]_S^{\leq}|/n$, so we have $RH_{R \cap S}(U) \geq RH_{R|S}(U)$.

Definition 11: Given U, R and S are two fuzzy ordinal relations induced by attributes B_1 and B_2. The fuzzy rank mutual information of B_1 and B_2 is defined as

$$RMI_{R,S}(U) = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{|[x_i]_R^{\leq}| \times |[x_i]_S^{\leq}|}{n \times |[x_i]_R^{\leq} \cap [x_i]_S^{\leq}|}. \tag{8}$$

Just like the relationship of information entropy, conditional entropy and mutual information in Shannon's information theory, we also have

$$RMI_{R,S}(U) = RH_R(U) - RH_{R|S}(U); \tag{9}$$

$$RMI_{R,S}(U) = RH_S(U) - RH_{S|R}(U). \tag{10}$$

The above definitions give a point-wise way to define rank entropy, rank conditional entropy and rank mutual information, and the entropy of the universe can be understood as the expectation of samples' entropy. Take $x_i \in U$ as an example.

$$\begin{aligned} & RMI_{R,S}(x_i) \\ & = -\log\left(|[x_i]_R^{\leq}| \times |[x_i]_S^{\leq}|\right) / \left(n \times |[x_i]_R^{\leq} \cap [x_i]_S^{\leq}|\right) \\ & = \log\left(n \times |[x_i]_R^{\leq} \cap [x_i]_S^{\leq}|\right) / \left(|[x_i]_R^{\leq}| \times |[x_i]_S^{\leq}|\right). \end{aligned}$$

For a set of given samples, n is a constant. Then we just consider $\theta = |[x_i]_R^{\leq} \cap [x_i]_S^{\leq}| / \left(|[x_i]_R^{\leq}| \times |[x_i]_S^{\leq}|\right)$. θ can be understood as a similarity function of two fuzzy ordinal sets. If $[x_i]_R^{\leq} = [x_i]_S^{\leq}$, $RMI_{R,S}(x_i) = -\log|[x_i]_S^{\leq}|/n = RH_S(x_i)$, which means the mutual information between R and S is equal to the rank entropy of R or S if they are the same.

Now we show some toy examples in Fig. 3. There are four classification tasks described with two features. The first one is a monotonically consistent classification task, while the second one is consistent but non-monotonic. The third one is monotonically consistent except two noisy samples, and the last one is a monotonic task with many inconsistent samples.

First, we consider (1) and (2). We compute the mutual information (MI) and rank mutual information (RMI) between features and decision of the two tasks. As to the first task, we obtain that MI=0.928, RMI=0.723; as to the second task, MI=0.952, RMI=0.472. We can see that mutual information does not vary much when the order of decision changes. It shows that mutual information is not sensitive to the ordinal structures of data. However, rank mutual information decreases from 0.723 to 0.472. It tells us the features in the second task are not good for monotonic classification. Compared with mutual information, rank mutual information reflects the ordinal structures of features.

Now, we consider (1) and (3). The dataset in (3) is generated from (1) by adding two noisy samples. We compute monotonic dependency and rank mutual information. The monotonic dependency of decision on the two features is 1 as the task is consistently monotonic, as shown in Fig. 3(1). In this case, RMI=0.723. However, if two noisy samples are added, shown in Fig. 3(3), monotonic dependency drops from 1 to 0.480, while RMI changes from 0.723 to 0.698. Obviously, the noisy samples have great impact on the monotonic dependency, while RMI is relatively robust. Fig. 3(4) presents a monotonic classification task. However, the class regions are overlapped. The rank mutual information RMI=0.665.

The above analysis shows rank mutual information can reflect the ordinal structures between attributes and decision, while mutual information fails it. Moreover, compared with monotonic dependency, rank mutual information is a robust metric for measuring monotonic consistency.

IV. Feature evaluation and selection

In this section, we discuss the technique to evaluate quality of features for monotonic classification based on rank mutual information. Then we combine the evaluation function with min-Redundancy-and-Max-Relevance (mRMR) [29] search strategy for selecting optimal features. Finally, two classification algorithms are introduced for validating selected features.

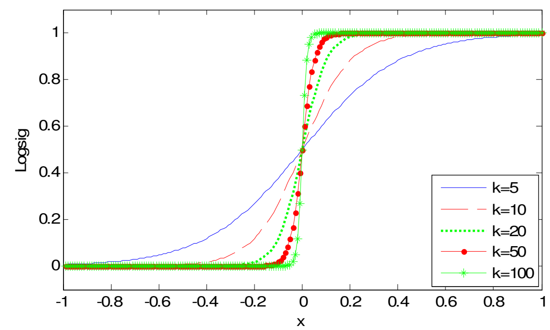


Fig. 4. Membership functions of fuzzy ordinal sets computed with logsig function under different parameter values

A. feature evaluation with RMI

In general classification learning, a set of learning samples is gathered and given to the learning algorithm. Let $< U, A, D >$ be the dataset. As to a monotonic classification task, we assume the values of D have the following relation: $d_1 < d_2 < \cdots < d_k$. Given attribute $a \in A$, we introduce the following relation function [10]

$$r_{ij}^{<} = \frac{1}{1 + e^{k(v(x_i,a)-v(x_j,a))}}, \tag{11}$$

where $v(x_i, a)$ is the value of x_i on feature a, and k is a positive constant. This function is the well-known 'logsig' transfer function used in neural networks. It is easy to see that $r_{ii}^{>} = r_{ii}^{<} = 0.5$, $r_{ij}^{<}$ approaches 1 if $v(x_j, a)$ is far larger than $v(x_i, a)$, and $r_{ij}^{<}$ approaches 0 if $v(x_j, a)$ is much smaller than $v(x_i, a)$. These results are consistent with our intuition: $r_{ij} = 0.5$ indicates that there is no difference between x_i and x_j; and $r_{ij}^{<} = 1$ indicates that x_i is much smaller than x_j and $r_{ij}^{<} = 0$ means x_i is much larger than x_j.

The relations computed with the above function are neither reflexive nor symmetric, but they are transitive, i.e., $r_{ii} \neq 1$, $r_{ij} \neq r_{ji}$ and $\min_y(R(x, y), R(y, z)) \leq R(x, z)$. The curves of the logsig function $f(x) = 1/(1 + \exp(-kx))$ under different values of k are shown in Fig. 4. Parameter k reflects user's preference and understanding of the words "larger" or "better". If k is very large, for example k=100, the fuzzy ordinal set can be understood as a fuzzy set of "slightly larger"; while if k is small, for example k=1, the corresponding fuzzy set is a set of "significantly larger".

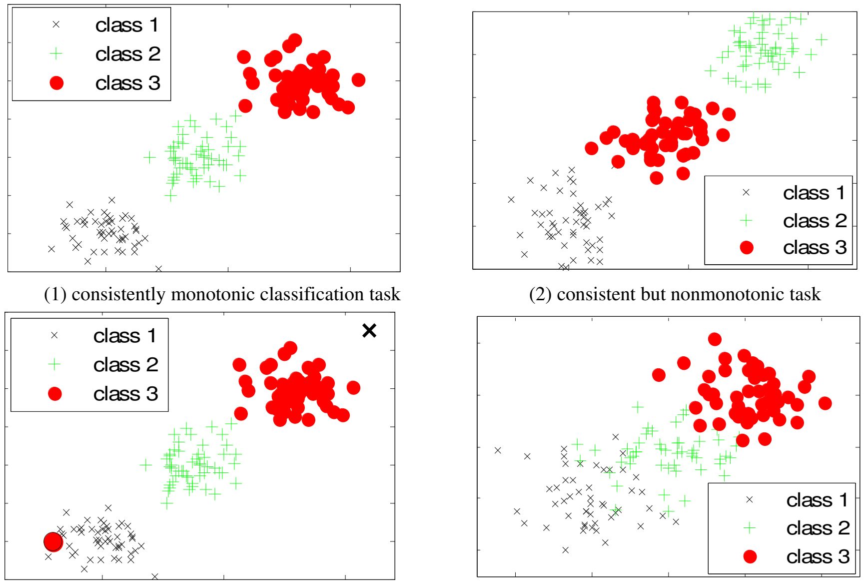
Now, we can get the fuzzy ordinal set which is larger than x_i in terms of attribute a, which is

$$[x_i]_a^{\leq} = r_{i1}/x_1 + r_{i2}/x_2 + \cdots + r_{in}/x_n. \tag{12}$$

Consider two features a_1 and a_2. If the fuzzy ordinal sets in terms of attributes a_1 and a_2 are $[x_i]_{a_1}^{\leq}$ and $[x_i]_{a_2}^{\leq}$, respectively, the fuzzy ordinal set in terms of $B = \{a_1, a_2\}$ is computed as $[x_i]_B^{\leq} = [x_i]_{a_1}^{\leq} \cap [x_i]_{a_2}^{\leq}$, where $\bigcap$ is a fuzzy intersection operator. Now the rank mutual information between a_1 and a_2 is calculated with

$$RMI_{a_1,a_2}(U) = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{|[x_i]_{a_1}^{\leq}| \times |[x_i]_{a_2}^{\leq}|}{n \times |[x_i]_{a_1}^{\leq} \cap [x_i]_{a_2}^{\leq}|}. \tag{13}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.



(1) consistently monotonic classification task

(2) consistent but nonmonotonic task

(3) monotonic task with two noisy samples

(4) monotonic task with overlapped regions

Fig. 3. Toy examples of classification tasks in 2-D feature spaces

Regarding features B and D, $RMI_{B,D}(U)$ reflects the monotonic consistency between B and D. We consider $RMI_{B,D}(U)$ as the significance of B in predicting D. For simplicity, $RMI_{B,D}(U)$ is also written as $RMI_{B,D}$ in the below.

Theorem 1: Given a monotonic classification dataset $< U, A, D >$, $B \subseteq A$. If D is completely dependent on B, we have $RMI_{B,D} = RH_D(U)$.

Proof: If D is completely dependent on B, the monotonic classification is consistent in this case. That is to say, $[x_i]_B^{\leq} \subseteq [x_i]_D^{\leq}$. Thus $RMI_{B,D} = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{|[x_i]_B^{\leq}|\times|[x_i]_D^{\leq}|}{n\times|[x_i]_B^{\leq}\cap[x_i]_D^{\leq}|} = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{|[x_i]_B^{\leq}|\times|[x_i]_D^{\leq}|}{n\times|[x_i]_B^{\leq}|} = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{|[x_i]_D^{\leq}|}{n} = RH_D(U)$.

Definition 12: Given a monotonic classification dataset $< U, A, D >$, $B \subseteq A$ and $a \in A - B$. The significance of a with respect to B is defined as $Sig(a, B, D) = RMI_{B\cup\{a\},D} - RMI_{B,D}$.

Corollary 1. Given a monotonic classification dataset $< U, A, D >$, $B \subseteq A$. If D is completely dependent on B and $a \in A - B$, we have $Sig(a, B, D) = 0$.

Proof: If D is completely dependent on B, $[x_i]_B^{\leq} \subseteq [x_i]_D^{\leq}$. Let $C = B\cup\{a\}$. $[x_i]_C^{\leq} = [x_i]_B^{\leq}\cap[x_i]_a^{\leq}$, thus $[x_i]_C^{\leq} \subseteq [x_i]_B^{\leq} \subseteq [x_i]_D^{\leq}$. $RMI_{B\cup\{a\},D} = RH_D(U)$ and $RMI_{B,D} = RH_D(U)$. $RMI_{B\cup\{a\},D} - RMI_{B,D}(U) = 0$. We obtain $Sig(a, B, D) = 0$.

Corollary 1 shows if features B have enough information for predicting D, any new feature cannot increase the mutual information. Thus the feature of significance zero can be removed from the system. By this way, the redundant or irrelevant features are reduced.

We can compute crisp or fuzzy ordinal relations from numerical features. If crisp relations are used, we call the mutual information as rank mutual information (RMI), while if fuzzy ordinal relations are considered, we call it fuzzy rank mutual information (FRMI).

B. min-Redundancy-and-Max-Relevance feature search

The above analysis introduced a feature evaluation metric for monotonic classification. Now we consider selecting features based on this metric.

A straightforward way is to exhaustively calculate the quality of feature subsets for finding an optimal subset. However, this is not feasible even given a moderate size of candidate features due to the exponential complexity.

Some efficient algorithms were developed to overcome this problem. Battiti in [28], and Peng, Long and Ding [29] discussed two criteria, named Max-Relevance (MR), min-Redundancy-and- Max-Relevance (mRMR), respectively. Intuitively, features of larger relevance with decision should provide more information for classification. Therefore, the best feature should be the one of the largest mutual information. This strategy is called maximal relevance criterion (Max-Relevance, MR). Formally, Max-Relevance criterion can be written as the following formulation:

$$max\Upsilon, \Upsilon = \frac{1}{|B|}\sum_{a_i\in B} RMI_{a_i,D}. \tag{14}$$

In essence the MR criterion is a feature selection algorithm based on ranking. We rank the features in the descending order according to the rank mutual information between single feature and decision, and then select the first k features, where k is specified in advance. It is well known that the ranking based algorithms cannot remove redundancy between features because this algorithm neglects the relevance between input variables. Sometimes, the redundancy between features is so great that deleting some of them does not reduce the classification information of the original data. In this case, we should select a subset of features with the minimal redundancy condition. That is

$$min(\Theta), \Theta = \frac{1}{|B|^2} \sum_{a_i, a_j \in B} RMI_{a_i, a_j}. \quad (15)$$

Then we get a new criterions $max\ \Phi(D, R)$, called minimal-Redundancy-Maximal-Relevance (mRMR), by combining the two constraints:

$$\Phi = \frac{1}{|B|} \sum_{a_i \in B} RMI_{a_i, D} - \frac{\beta}{|B|^2} \sum_{a_i, a_j \in B} RMI_{a_i, a_j}, \quad (16)$$

where the parameter β is used to regulate the relative importance of the mutual information between features and decision.

In [29], an incremental version of mRMR was developed. If a subset B of $l-1$ features has been selected in current step, now we select the $l'th$ feature. The incremental algorithm computes the following metric. $\forall a_j \in A - B$,

$$Sig(a_j, B, D) = RMI_{a_j, D} - \frac{\beta}{l-1} \sum_{a_i \in B} RMI_{a_i, a_j}. \quad (17)$$

The feature a maximizing $Sig(a; B, D)$ is selected.

mRMR calculates the significance of each feature one by one, and finally we get the rank of the features. Then some classification algorithm should be introduced to check the best k features with respect to the classification performance via cross validation.

In the incremental algorithm, we should compute the mutual information between the remaining $m-(l-1)$ features and decision attribute. Moreover, we also require calculating the mutual information between the remaining $m-(l-1)$ and the selected $l-1$ features. Thus the total computational cost is $m-(l-1)+(m-(l-1)) \times (l-1)= (m-l+1) \times l$ in this step. In fact, this algorithm just uses the mutual information between features pairs, and the mutual information between features and decision. So we can compute and store the matrix of mutual information M_{ij} in advance, where M_{ij} is the mutual information between feature pairs. In this case, the total computational cost is $m+m \times m$, where m for computing mutual information between m single features and decision, while $m \times m$ for computing mutual information between feature pairs.

V. Experimental analysis

In this section, we present some experiments on real-world tasks to test the proposed technique. We compare our metrics with the dependency functions defined in dominance rough sets and fuzzy preference rough sets to show the robustness of rank mutual information. And we also compare rank mutual information with mutual information and fuzzy mutual information for showing the effectiveness of these metrics in measuring monotonic consistency.

We introduce two monotonic classifiers, OLM [12] and OSDL [4], [23], to calculate the classification performance of the selected features. These algorithms are now implemented in Weka [51].

Here mean absolute error (MAE) is introduced to evaluate decision performance, computed as

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|, \quad (18)$$

TABLE I
Data description

Data set	Instances	Features	Classes
Adult	48842	14	2
Ailerons	13750	41	3
Auto MPG	398	8	3
Australian Credit	690	15	2
Bankruptcyrisk	36	12	3
Cardiotocography	2126	21	3
Credit Approval	690	14	2
Fault	540	52	5
German Credit	1000	20	2
Housing	506	13	4
Pasture	36	22	3
Triazines	186	61	3
Windsor Housing	546	11	4
Wine Quality-red	1599	11	6

where N is the number of samples in the test set and $\widehat{y_i}$ is the output of the algorithm and y_i is the real output of the i'th sample. Moreover, we also compute the classification error rate(CE) of models.

Fourteen monotonic tasks are collected from UCI machine learning repository and other webpages [57]. The detailed information about these datasets is given in Table 1.

We randomly select three tasks, including Adult, Pasture and Wine Quality-red, to test the robustness of rank mutual information and monotonic dependency defined in dominance rough sets and fuzzy preference rough sets. We calculate dependency or mutual information between decision and each single feature based on the raw datasets. Moreover, we randomly draw $k\%$(k=5, 10, 15, 20) samples from the raw datasets and replace their class labels with arbitrary candidates. These samples are considered to be class-noisy and are put back to the raw datasets. Now we observe the variation of dependency or mutual information.

If the metric is robust, we expect the value variation of metric will be small. Thus the difference of metrics computed at different levels of noise would be small enough. Monotonic dependency (PRS) and monotonic fuzzy dependency (FPRS), rank mutual information (RMI) and fuzzy rank mutual information (FRMI) computed with the raw datasets and noisy datasets are shown in Fig. 5 to Fig. 7.

Observing the curves Fig. 5, we see that PRS changes much at different levels of noise. The metric values are completely different from the value computed with the raw dataset when 5% noisy samples are added in Adult. As to FPRS, although the metric values also vary, it seems that this metric is more robust than PRS. RMI and FRMI are far more robust than PRS and FPRS as the metric values are stable when noisy samples are added. The same conclusions can also be drawn from Fig. 6 and Fig. 7.

Now we compare the classification performance of different feature subsets selected with 6 metric functions: mutual information (MI), fuzzy mutual information (FMI), rank mutual information (RMI), fuzzy rank mutual information (FRMI), PRS and FPRS. We rank the features with the descending order of these metric values and added the top features

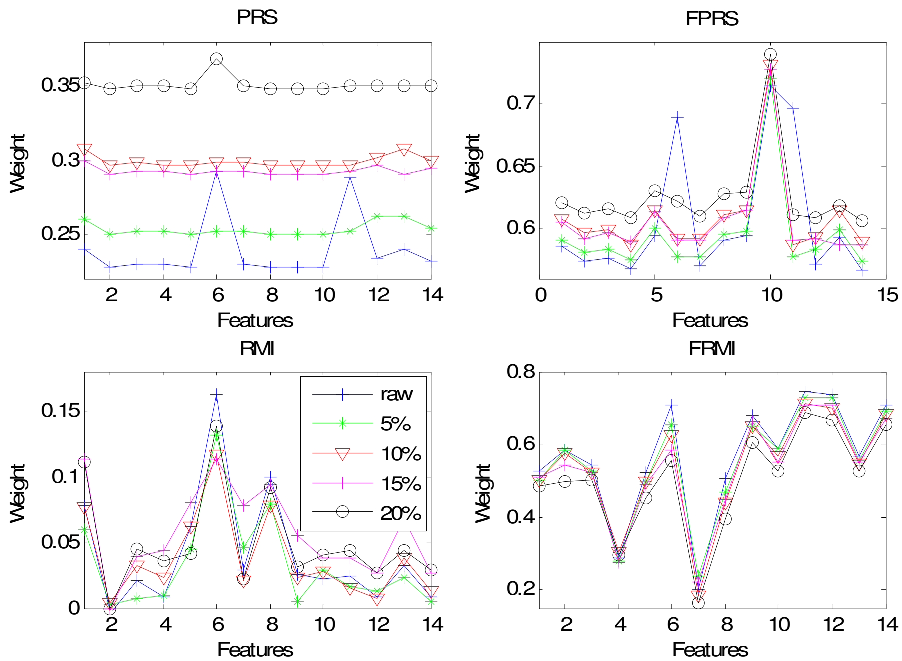


Fig. 5. Metric of monotonic consistency computed at different noise levels (Adult)

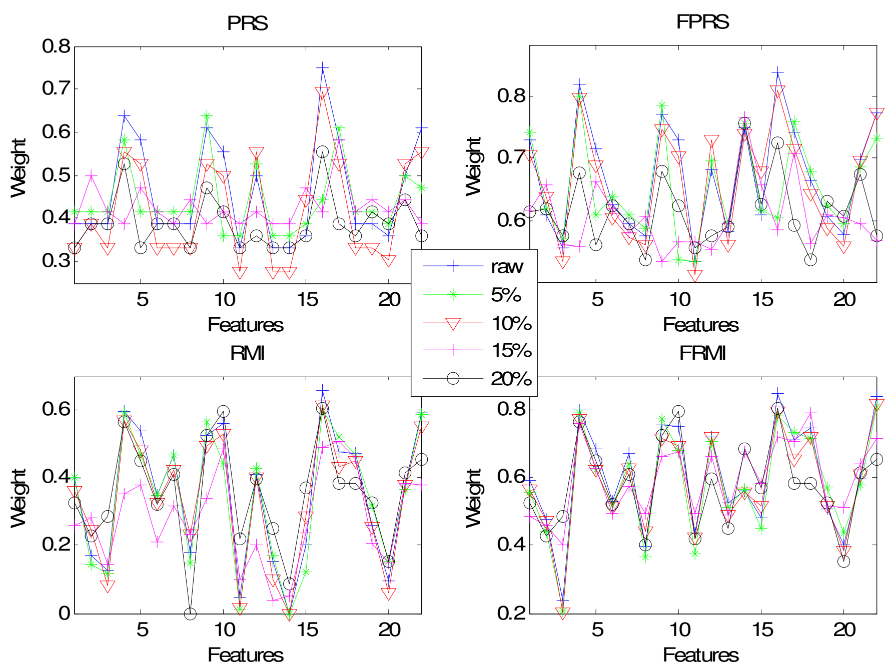


Fig. 6. Metric of monotonic consistency computed at different noise levels (Pasture)

one by one. In this process, we compute the classification performances of the corresponding features based on 5-fold cross validation. Two metrics of classification performance are calculated: classification error rate (CE) and mean absolute error (MAE). The best classification performance and the number of the corresponding subsets of features are given in Tables 2 and 3, where raw is the performance of the raw datasets. Due to the limitation of space, here we just give the average performance of cross validation in these tables.

We first compare the performances computed before and after feature selection. We see that both the classification error rate and mean absolute error decrease after some features are reduced from the raw datasets. As to some tasks like Bankruptcyrisk and Cardiotocography, feature selection significantly improves the classification.

RMI, FRMI, PRS and FPRS are the evaluation functions considering the ordinal structures of features and decision, while MI and FMI do not reflect the monotonic consistency.

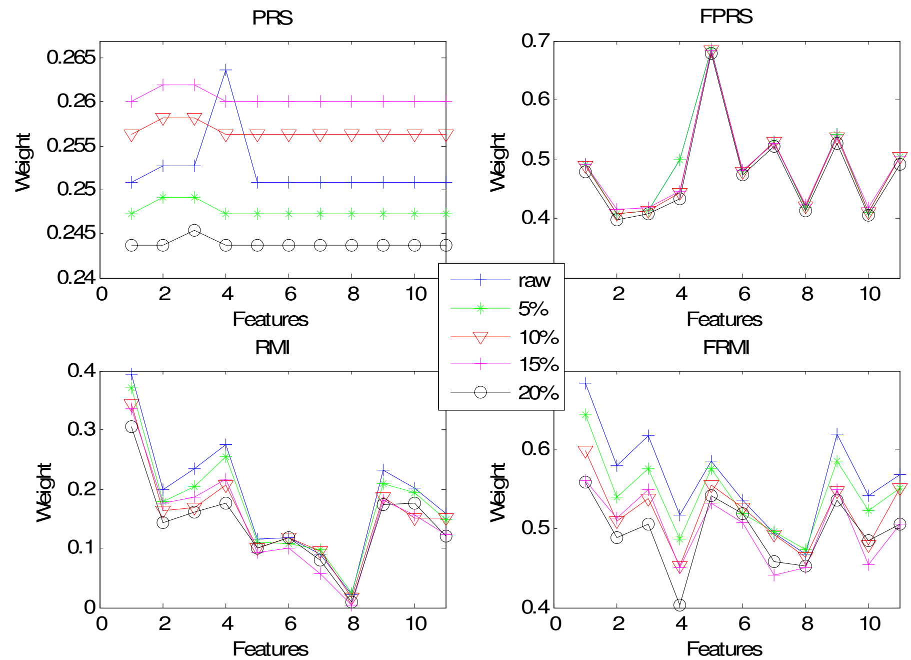


Fig. 7. Metric of monotonic consistency computed at different noise levels (Wine Quality-red)

We can see that most tasks produce the better classification performances after attribute reduction based on RMI, FRMI, PRS and FPRS as to OLM and OSDL. Although MI and FMI show good performances in feature evaluation for general classification tasks, they are worse than RMI and FRMI, even worse than PRS and FPRS. In addition, RMI and FRMI are better than PRS and FPRS in most cases.

Fig. 8 and Fig. 9 present the curves of error rate varying with the number of the selected features. We consider six tasks in these figures. As to OLM, the classification error rates decrease when the first several features are used, and then the error rates increase after they arrive at their minimums. This trend is the same as those in feature selection for general classification tasks [58], [59]. The first several features are useful for classification learning. However, too many selected features may lead to the issue of overfitting. Thus a proper number of features are very important for obtaining good classification performance.

As to OSDL, no consistent rule can be drawn from these curves. Although feature reduction improves classification performance of OSDL, these error rates do not decrease with the addition of new features. It is easy to select the best features for OSDL.

Rank based feature selection does not consider the redundant information between the selected features, which may result in superfluous features. The strategy of Min redundancy and Maximal relevance (mRMR) consider not only the relevance between decision and features, but also the relevance between features. The features high relevant to decision and low correlated with the selected features are considered to be useful.

Now we compare the features selected with mRMR strategy, where MI, FMI, RMI and FRMI are all employed to evaluate the candidate features. As mentioned above, mRMR just outputs a rank of candidate features. We should introduce other classification algorithms to validate the feature subsets. Here we also consider OLM and OSDL. The performances of the best subsets of features are given in Table 4.

Comparing the results in Table 4, we see that combination of RMI or FRMI with mRMR is much better than integration of MI or FMI with mRMR. Regarding OLM, RMI and FRMI are better than MI and FMI on twelve of fourteen tasks. As to OSDL, RMI and FRMI outperform MI and FMI on eleven of fourteen tasks, and they obtain the same performance on all the other datasets, which show RMI and FRMI are no worse than MI and FMI. In addition, as to OSDL, it is notable that FRMI based mRMR obtains the best performances on nine of the tasks, which reflects this attribute reduction algorithm is effective for OSDL.

VI. Conclusions and future work

Monotonic classification is a class of special tasks in machine learning. Monotonicity constraints are considered as the fundamental assumption about these tasks. However, existing techniques for classification modeling are not applicable to this domain because they fail to discover and represent the ordinal structures of datasets. In this work, we introduce a metric function, called rank mutual information, to measure monotonic consistency between features and decision. We then combine this function with the min-Redundancy-and-Max-Relevance search strategy for constructing an effective feature selection algorithm for monotonic classification. Some numerical experiments are conducted to test the performance of the proposed algorithm. The following conclusions are drawn from the analysis.

TABLE II
OLM PERFORMANCE OF THE SUBSETS OF FEATURES SELECTED WITH DIFFERENT EVALUATION FUNCTIONS (%)

Data set		MI	RMI	FMI	FRMI	PRS	FPRS	raw
Adult	CE	25.0±1.6(12)	25.0±1.6(3)	26.8±3.0(8)	25.6±2.9(14)	28.6±4.0(13)	25.6±2.9(10)	28.6±4.8
	MAE	25.0±1.6(12)	25.0±1.6(3)	26.8±3.0(8)	25.6±2.9(14)	28.6±4.0(13)	25.6±2.9(10)	28.6±4.8
Ailerons	CE	53.3±3.5(2)	37.0±3.8(19)	55.2±5.0(4)	36.9±3.9(14)	36.3±3.4(37)	32.7±3.0(15)	61.9±7.7
	MAE	35.6±5.3(2)	24.7±3.1(19)	36.8±4.5(4)	24.6±4.5(14)	24.2±4.9(37)	21.8±3.2(15)	41.3±13.9
Auto MPG	CE	29.3±4.0 (5)	28.6±3.1(5)	28.6±3.1(5)	28.6±3.1(5)	28.6±3.1(5)	28.3±2.8(7)	32.0±5.1
	MAE	19.6±7.0(5)	19.1±4.6(5)	19.1±4.6(5)	19.1±4.6(5)	18.9±4.6(7)	17.0±6.5(4)	22.0±6.8
Australian Credit	CE	16.5±4.9(6)	16.5±4.9(6)	17.0±4.9(11)	15.1±3.5(5)	17.1±1.4(10)	17.1 ±1.4(10)	23.6±2.8
	MAE	16.5±4.9(6)	16.5±4.9(6)	17.0±4.9(11)	15.1±3.5(5)	17.1±1.4(10)	17.1 ±1.4(10)	23.6±2.8
Bankruptcyrisk	CE	17.9 ±6.6(2)	17.9±6.6(2)	17.9±6.6(2)	17.9±6.6(2)	17.9±6.6(2)	17.9±6.6(2)	51.3±18.0
	MAE	12.0±6.6(2)	12.0±6.6(2)	12.0±6.6(2)	12.0±6.6(2)	12.0±6.6(2)	12.0±6.6(2)	34.2±18.0
Cardiotocography	CE	10.8±0.7(12)	10.5±1.1(16)	10.6±1.3(17)	10.5±1.1(16)	11.9±2.9(12)	10.6±1.2(15)	12.3±0.64
	MAE	7.2±0.4(12)	7.0±1.6(16)	7.1±1.7(17)	7.0±1.6(16)	7.9±3.2(12)	7.1±1.4(15)	8.2±1.4
Credit Approval	CE	17.0±3.6(6)	11.3±2.2(3)	17.4±4.0(7)	11.2±2.2(1)	20.3±2.6(12)	19.0±4.3(10)	25.8±6.1
	MAE	17.0±3.6(6)	11.3±2.2(3)	17.4±4.0(7)	11.2±2.2(1)	20.3±2.6(12)	19.0±4.3(10)	25.8±6.1
Fault	CE	3.7±2.2(49)	3.7±2.2(49)	3.7±2.2(47)	3.5±2.0(49)	3.3±1.9(44)	3.7±1.9(51)	3.9±1.8
	MAE	1.5±5.8(49)	1.5±5.8(49)	1.5±6.5(47)	1.4±5.7(49)	1.3±5.7(44)	1.5±5.3(51)	1.6±5.2
German Credit	CE	30.0±0.0(1)	29.0±1.9(4)	30.0±0.0(1)	27.8±2.6(7)	30.0±0.0(1)	29.9±0.0(1)	36.0±3.0
	MAE	30.0±0.0(1)	29.0±1.9(4)	30.0±0.0(1)	27.8±2.6(7)	30.0±0.0(1)	29.9±0.0(1)	36.0±3.0
Housing	CE	32.4±3.3(12)	32.2±2.7(10)	33.0±2.6(6)	33.0±3.6(10)	32.2±2.7(10)	33.8±3.7(12)	35.4±4.4
	MAE	16.2±5.1(12)	16.1±6.5(10)	16.5±4.0(10)	16.5±4.7(10)	16.1±4.4(10)	16.9±5.5(12)	17.7±7.3
Pasture	CE	11.1±10.7(2)	13.9±10.7(11)	11.2±10.7(2)	11.1±10.7(2)	16.7±13.5(4)	16.7±13.5(4)	22.2±8.3
	MAE	7.4±10.7(2)	9.3±0.8(11)	7.4±10.7(2)	7.4±10.7(2)	11.1±13.5(4)	11.1±13.5(4)	14.8±12.4
Triazines	CE	55.4±1.9(2)	50.0±3.3(2)	54.8±5.7(5)	51.6±11.9(45)	45.7±10.5(25)	51.1±7.8(26)	61.3±9.0
	MAE	36.9±8.2(2)	33.3±4.3(2)	36.6±11.1(5)	34.4±17.3(45)	30.5±11.6(25)	34.1±10.6(26)	40.9±17.1
Windsor Housing	CE	48.2±6.6(10)	47.4±7.3(10)	48.2±6.6(10)	47.4±6.3(11)	48.4±8.1(11)	46.7±8.1(9)	49.1±7.2
	MAE	24.1±9.0(10)	23.7±9.2(10)	24.1±9.0(10)	23.7±7.6 (11)	24.2±9.5(11)	23.4±10.3(11)	25.5±9.0
Wine Quality-red	CE	47.5±2.6(1)	47.5±2.6(1)	47.0±2.7(7)	47.5±2.6(1)	51.2±2.2(11)	50.4±3.0(6)	52.6±2.7
	MAE	15.8±2.3(1)	15.8±2.3(1)	15.7±3.5(7)	15.8±2.3(1)	17.1±2.7(11)	16.8±3.7(6)	17.5±5.2

TABLE III
OSDL PERFORMANCE OF SUBSETS OF FEATURES SELECTED WITH DIFFERENT EVALUATION FUNCTIONS (%)

Data set		MI	RMI	FMI	FRMI	PRS	FPRS	raw
Adult	CE	22.8±0.5(1)	21.6±3.3(3)	22.8±0.5(1)	21.6±3.3(3)	22.8±0.5(1)	22.8±0.5(1)	26.6±2.1
	MAE	22.8±0.5(1)	21.6±3.3(3)	22.8±0.5(1)	21.6±3.3(3)	22.8±0.5(1)	22.8±0.5(1)	26.6±2.1
Ailerons	CE	39.1±2.6(1)	33.0±4.2(4)	39.1±2.6(1)	32.6±4.2(4)	33.4±4.4(38)	32.7±5.5(15)	54.1±7.5
	MAE	26.1±2.9(1)	22.0±4.8(4)	26.1±2.9(1)	21.7±4.8(4)	22.3±4.5(38)	21.8±5.8(15)	36.1±7.7
Auto MPG	CE	25.3±2.4(2)	24.8±3.2(4)	27.0±2.9(4)	27.0±2.9(3)	27.0±2.4(4)	25.5±2.4(3)	33.9±1.4
	MAE	16.8±3.7(2)	16.2±3.1(4)	18.0±3.5(4)	18.0±3.9(3)	18.0±3.6(4)	17.0±3.6(4)	22.6±4.0
Australian Credit	CE	14.4±3.9(6)	14.4±3.9(6)	14.4±3.9(6)	13.5±3.5(1)	20.3±3.3(12)	0.1±0.3(3)	20.3±4.1
	MAE	14.4±3.9(6)	14.4±3.9(6)	14.4±3.9(6)	13.5±3.5(1)	20.3±3.3(12)	0.1±0.3(3)	20.3±4.1
Bankruptcyrisk	CE	12.8±13.5(6)	7.7±11.4(3)	12.8±13.5(6)	7.7±11.4(3)	10.3±10.6(6)	7.7±11.4(3)	30.8±10.9
	MAE	8.6±13.5(6)	5.1±11.4(3)	8.6±13.5(6)	5.1±11.4(3)	6.8±10.6(6)	5.1±11.4(3)	20.5±10.9
Cardiotocography	CE	13.5±1.7(3)	12.2±1.0(4)	12.1±1.8(3)	10.9±1.7(4)	11.1±4.8(4)	10.6±1.2(15)	27.8±2.6
	MAE	9.0±2.1(3)	8.2±1.4(4)	8.1±2.3(3)	7.2±2.1(4)	7.4±5.0(4)	7.1±1.4(15)	18.5±2.7
Credit Approval	CE	11.9±2.2(1)	11.9±2.2(1)	11.9±2.2(1)	11.9±2.2(1)	13.5±0.7(9)	11.9±2.2(1)	26.4±4.3
	MAE	11.9±2.2(1)	11.9±2.2(1)	11.9±2.2(1)	11.9±2.2(1)	13.5±0.7(9)	11.9±2.2(1)	26.4±4.3
Fault	CE	36.5±3.9(17)	36.5±3.9(17)	38.5±6.0(21)	37.4±2.7(18)	42.6±5.2(27)	43.7±3.0(32)	63.0±2.9
	MAE	14.6±4.6(17)	14.6±4.6(17)	15.4±9.0(21)	15.0±4.9(18)	17.0±7.0(27)	17.5±1.8(32)	25.2±5.9
German Credit	CE	28.3±1.4(1)	28.1±1.6(4)	28.3±1.4(1)	28.1±1.6(6)	30.0±0.0(1)	29.5±0.8(5)	50.5±2.2
	MAE	28.3±1.4(1)	28.1±1.6(4)	28.3±1.4(1)	28.1±1.6(6)	30.0±0.0(1)	29.5±0.8(5)	50.5±2.2
Housing	CE	31.2±4.4(7)	32.4±3.9(9)	32.2±3.7(7)	34.0±3.3(10)	33.4±3.9(4)	34.0±3.3(12)	36.0±3.1
	MAE	15.6±5.8(7)	16.2±4.1(9)	16.1±3.3(7)	17.0±2.7(10)	16.7±4.7(4)	17.0±2.0(12)	18.0±2.2
Pasture	CE	25.0±10.1(1)	19.4±13.0(5)	25.0±10.1(2)	22.2±14.8(4)	19.4±8.2(5)	19.4±8.2(6)	63.9±12.4
	MAE	16.7±10.1(1)	13.0±13.0(5)	16.7±10.1(2)	14.8±14.8(4)	13.0±8.2(5)	13.0±8.2(6)	42.6±12.4
Triazines	CE	51.6±7.4(30)	48.4±4.1(33)	55.9±6.7(13)	46.8±9.9(34)	45.7±5.8(22)	49.5±4.9(60)	55.9±4.9
	MAE	34.4±10.1(30)	32.3±5.2(33)	37.3±10.1(13)	31.2±15.2(34)	30.5±7.1(22)	33.0±5.5(60)	37.3±5.5
Windsor Housing	CE	45.1±7.1(11)	45.1±7.1(11)	45.1±7.1(11)	45.1±7.1(11)	45.1±7.1(11)	42.9±6.8(9)	45.1±7.1
	MAE	22.5±8.1(11)	22.5±8.1(11)	22.5±8.1(11)	22.5±8.1(11)	22.5±8.1(11)	21.4±8.5(9)	22.5±8.1
Wine Quality-red	CE	46.4±1.1(11)	43.9±1.0(2)	48.2±3.9(3)	44.2±1.6(2)	47.2±3.2(2)	44.6±1.1(5)	55.3±3.2
	MAE	15.5±1.2(1)	14.6±1.6(2)	16.1±4.5(3)	14.7±2.1(2)	15.7±3.4(2)	14.9±1.5(5)	18.5±6.0

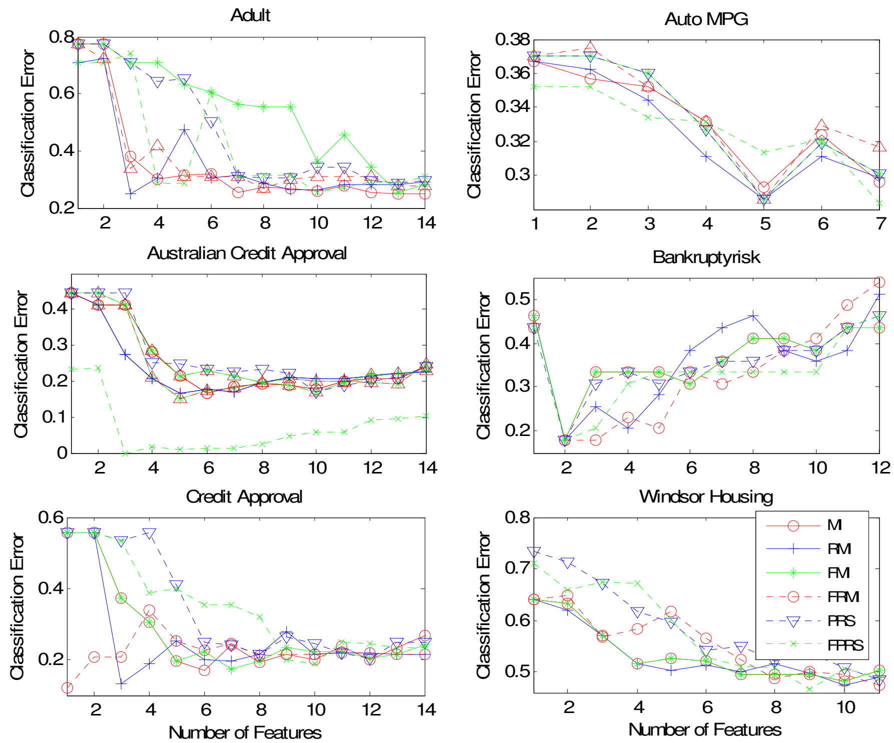


Fig. 8. OLM performance curves with number of features

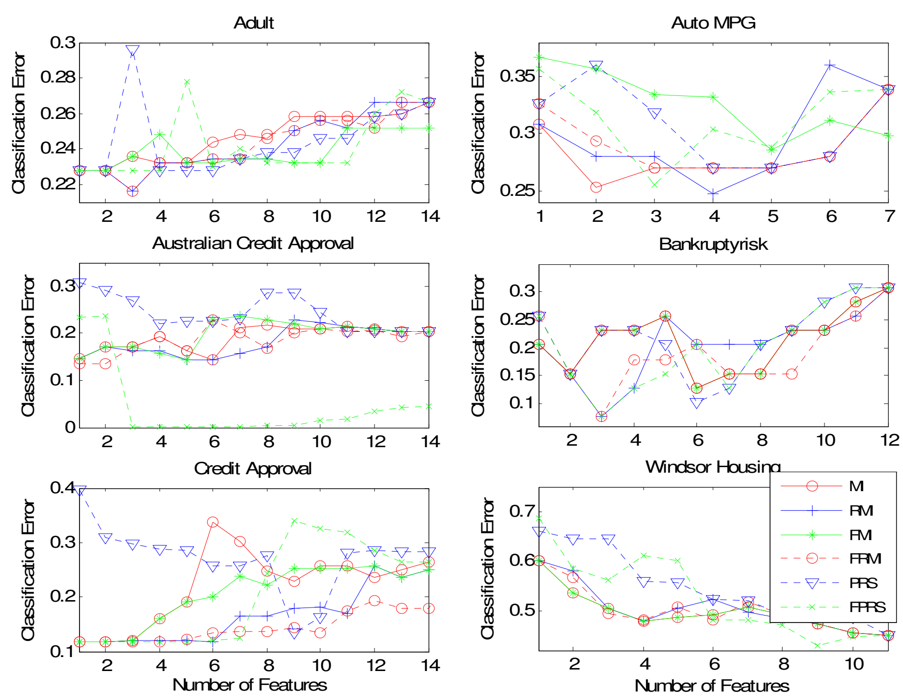


Fig. 9. OSDL performance curves with number of features

First, although mutual information and fuzzy mutual information are effective and robust functions in measuring feature quality for general classification tasks, they are not applicable to monotonic tasks because they can not able to reflect the ordinal structures.

Second, monotonic dependency functions defined in dominance rough sets and fuzzy preference rough sets are able to reflect the monotonic relevance between features and decision. However, these functions are very sensitive to noisy samples. It makes them not applicable in noisy conditions.

Finally, rank mutual information and fuzzy rank mutual information combine the advantages of mutual information and

TABLE IV
mRMR BASED FEATURE SELECTION, WHERE FEATURES ARE EVALUATED WITH MI, FMI, RMI AND FRMI (%)

Data set		OLM				OSDL			
		MI	RMI	FMI	FRMI	MI	RMI	FMI	FRMI
Adult	CE	29.6±3.8(14)	28.6±6.4(13)	30.8±3.8(14)	29.8±8.3(12)	22.8±0.6(1)	22.8±0.6(1)	22.8±0.6(1)	22.8±0.6(1)
	MAE	29.6±3.8(14)	28.6±6.4(13)	30.8±3.8(14)	29.8±8.3(12)	22.8±0.6(1)	22.8±0.6(1)	22.8±0.6(1)	22.8±0.6(1)
Ailerons	CE	53.7±3.2(4)	35.8±1.5(17)	55.2±5.9(4)	36.6±3.9(15)	39.1±2.6(1)	31.4±5.0(12)	39.1±5.0(1)	31.6±4.7(4)
	MAE	35.8±3.6(4)	23.9±1.9(17)	36.8±7.3(4)	24.4±3.3(15)	26.1±2.9(1)	20.1±5.2(12)	26.1±5.2(1)	21.1±5.0(4)
Auto MPG	CE	29.1±2.5(7)	28.6±4.2(7)	29.3±3.1(7)	28.8±1.3(7)	28.8±5.1(5)	28.6±3.1(5)	30.9±3.7(1)	28.1±4.3(6)
	MAE	17.4±6.3(7)	19.1±5.2(7)	19.6±3.0(7)	19.2±1.9(4)	19.5±10.4(1)	19.1±4.3(5)	20.6±4.5(1)	18.7±4.4(6)
Australian Credit	CE	19.0±2.7(7)	16.5±3.5(5)	19.0±2.7(7)	17.5±3.6(5)	14.5±3.2(1)	14.1±3.7(5)	14.2±3.4(3)	14.1±3.7(5)
	MAE	19.0±2.7(7)	16.5±3.5(5)	19.0±2.7(7)	17.5±3.6(5)	14.5±3.2(1)	14.1±3.7(5)	14.2±3.4(3)	14.1±3.7(5)
Bankruptcyrisk	CE	17.9±6.6(2)	17.9±6.6(2)	17.9±6.6(2)	17.9±6.6(2)	7.7±7.2(4)	7.7±7.2(4)	7.7±7.2(4)	7.7±7.2(4)
	MAE	12.0±6.6(2)	12.0±6.6(2)	12.0±6.6(2)	12.0±6.6(2)	5.1±7.2(4)	5.1±7.2(4)	5.1±7.2(4)	5.1±7.2(4)
Cardiotocography	CE	10.8±1.2(15)	10.2±1.8(15)	11.0±1.3(18)	10.7±1.4(10)	17.8±1.1(4)	16.4±0.7(2)	10.9±1.3(8)	10.5±2.8(2)
	MAE	7.2±1.7(15)	6.8±2.4(15)	7.3±1.2(18)	7.2±1.6(10)	11.9±1.5(4)	10.9±1.0(2)	7.3±1.4(8)	7.0±2.7(2)
Credit Approval	CE	18.7±2.5(6)	18.6±1.2(12)	21.0±4.0(13)	21.0±2.3(19)	11.9±2.2(1)	11.7±2.3(3)	11.9±2.2(1)	11.7±2.3(3)
	MAE	18.7±2.5(6)	18.6±1.2(12)	21.0±4.0(13)	21.0±2.3(19)	11.9±2.2(1)	11.7±2.3(3)	11.9±2.2(1)	11.7±2.3(3)
Fault	CE	10.4±2.7(51)	3.9±1.8(50)	3.7±2.0(49)	3.5±1.8(49)	36.7±4.6(9)	25.9±4.6(7)	38.3±5.3(12)	37.4±3.2(18)
	MAE	4.2±7.5(51)	1.6±4.6(50)	1.5±5.2(49)	1.4±5.1(49)	14.7±6.9(9)	10.4±6.5(7)	15.3±6.3(12)	15.0±4.9(18)
German Credit	CE	29.9±0.22(3)	29.2±2.4(5)	29.8±2.2(7)	28.2±2.0(6)	29.8±2.2(4)	28.4±1.7(4)	30.0±0.0(1)	28.1±3.3(5)
	MAE	29.9±0.22(3)	29.2±2.4(5)	29.8±2.2(7)	28.2±2.0(6)	29.8±2.2(4)	28.4±1.7(4)	30.0±0.0(1)	28.1±3.3(5)
Housing	CE	32.8±3.8(13)	30.4±4.4(7)	35.0±6.1(13)	33.2±2.0(9)	34.0±5.5(8)	33.4±4.6(8)	36.0±3.1(13)	34.4±2.9(9)
	MAE	16.4±6.7(13)	15.2±6.4(7)	17.5±8.5(13)	16.6±4.1(9)	17.0±4.8(8)	16.7±4.8(8)	18.0±2.2(13)	17.2±3.0(9)
Pasture	CE	22.2±7.3(5)	16.712.1(3)	27.8±10.2(5)	16.7±23.5(3)	22.2±16.3(4)	16.7±23.5(3)	22.2±16.3(4)	16.7±23.5(3)
	MAE	14.8±6.2(5)	11.1±12.1(3)	18.5±16.7(5)	11.1±23.5(3)	14.8±16.3(4)	11.1±23.5(3)	14.8±16.3(4)	11.1±23.5(3)
Triazines	CE	53.2±7.0(32)	46.2±4.7(5)	51.1±4.4(3)	47.3±9.0(5)	51.1±4.7(27)	49.5±4.4(6)	51.6±10.2(9)	50.0±6.6(12)
	MAE	35.5±14.9(32)	30.1±6.4(5)	34.1±5.7(3)	31.5±16.3(5)	34.1±9.8(27)	33.0±6.5(6)	34.4±19.0(9)	33.3±7.7(12)
Windsor Housing	CE	49.9±8.2(11)	49.1±5.2(8)	49.9±8.2(11)	46.9±7.7(10)	45.1±7.1(11)	45.1±7.1(11)	45.1±7.1(11)	45.1±7.1(11)
	MAE	24.6±11.2(11)	24.5±7.4(8)	24.6±5.7(11)	23.4±10.7(10)	22.5±8.1(11)	22.5±8.1(11)	22.5±8.1(11)	22.5±8.1(11)
Wine Quality-red	CE	46.5±2.6(2)	50.5±3.0(2)	50.3±1.9(1)	47.7±3.4(2)	40.0±1.3(4)	40.0±1.3(4)	39.9±0.8(5)	39.1±2.8(6)
	MAE	15.5±2.2(2)	16.8±4.9(2)	16.8±2.2(11)	15.9±2.6(2)	13.3±1.1(4)	13.3±1.1(4)	13.3±0.9(5)	13.0±2.5(6)

dominance rough sets. They are robust to noisy information and effective in reflecting ordinal structures. The proposed feature selection algorithms are competent with mutual information based mRMR algorithm in monotonic classification.

In this work, we talk about classification tasks which take assumption that all features are monotonic with decision. In real world applications, this assumption may not be true. Sometimes, just some features, instead of all, have monotonic relations with decision. Those non-monotonic features are also useful for constructing accurate predicting models. In decision analysis, monotonic features are called criteria, while non-monotonic features are called attributes. The corresponding task is called multi-criteria and multi-attribute decision analysis (MCMA). No much attention has been paid to this problem so far. We will develop feature selection techniques for it in future.

REFERENCES

[1] R. Potharst, A. J. Feelders, *Classification trees for problems with monotonicity constraints*, ACM SIGKDD Explorations Newsletter. vol. 4, no. 1, pp.1-10, 2002.

[2] J. Wallenius, J. S. Dyer, P. C. Sishburn, R. E. Steuer, S. Zionts, K. Deb, *Multiple criteria decision making, Multiattribute utility theory: recent accomplishments and what lies ahead*, Management Science. Vol. 54, no. 7, pp. 1336-1349, 2008.

[3] V. Popova, *Knowledge Discovery and Monotonicity*, PhD thesis, Erasmus University Rotterdam, 2004.

[4] C.-V. Kim, *Supervised ranking from semantics to algorithms*, PhD thesis, Ghent University, Belgium, 2003.

[5] W. Kotllowski, *Statistical Approach to Ordinal Classification with Monotonicity Constraints*, PhD thesis, Poznan University of Technology, 2008.

[6] S. Greco, B. Matarazzo, R. Slowinski, *Rough approximation of a preference relation by dominance relations*, European Journal of Operational Research 1999, vol. 117, pp. 63-83.

[7] S. Greco, B. Matarazzo, R. Slowinski, *Rough approximation by dominance relations*, Int. J. of Intell. Syst. 2002, vol. 17, pp. 153-171.

[8] Y. Sai, Y. Y. Yao, *N. Zhong. Data analysis and mining in ordered information tables*, In: Proceedings of the IEEE international conference on data mining 2001, pp. 497-504.

[9] W. Kotlowski, K. Dembczynski, S. Greco, R. Slowinski, *Stochastic dominance-based rough set model for ordinal classification*, Inf. Sci., 178, pp. 3989-4204, 2008.

[10] Q. H. Hu, D. R. Yu, M. Z. Guo, *Fuzzy preference based rough sets*, Information Sciences, vol. 180, no. 10, pp. 2003-2022, 2010.

[11] Q. H. Hu, M. Z. Guo, D. R. Yu, J. F. Liu, *Information entropy for ordinal classification*, Science in China Series F: Information sciences. vol. 53, no. 6, pp. 1188-1200, 2010.

[12] A. Ben-David, L. Sterling, Y. H. Pao, *Learning and classification of monotonic ordinal concepts*, Comput. Intell. 1989, vol. 5, pp. 45-49.

[13] A. Ben-David, *Monotonicity maintenance in information-theoretic machine learning algorithms*, Mach. Learn. 1995, vol. 19, pp. 29-43.

[14] W. Duivesteijn, *A. Feelders. Nearest Neighbour Classification with Monotonicity Constraints*, W. Daelemans et al. (Eds.): ECML PKDD 2008, Part I, LNAI 5211, pp. 301-316, 2008, Springer-Verlag Berlin Heidelberg 2008.

[15] N. Barile, A. Feelders, *Nonparametric Monotone Classification with MOCA*, Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on, pp. 731- 736, 2008.

[16] R. Potharst, J. C. Bioch, *Decision trees for ordinal classification*, Intelligent Data Analysis 4(2), 97-112, 2000.

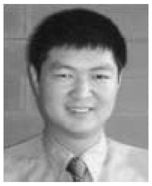
[17] K. Cao-Van, B. D. Baets, *Growing decision trees in an ordinal setting*, Int. J. of Intell. Syst., 2003, 18: 733-750.

[18] H. A. M. Daniels, M. V. Velikova, *Derivation of Monotone Decision Models From Noisy Data*, IEEE Trans. Syst., Man, Cybern. C, : Appl. and Rev., vol. 36, no. 5, pp. 705-710, 2006.

[19] A. J. Feelders, Martijn Pardoel, *Pruning for Monotone Classification Trees*, Lecture Notes in Computer Science, 2003, vol. 2811, pp. 1-12.
[20] R. V. Kamp, A. J. Feelders, N. Barile, *Isotonic Classification Trees*, In N. Adams (Ed.), Proceedings of IDA 2009, pp. 405-416.
[21] A. Jimnez, F. Berzal, J.-C. Cubero, *POTMiner: mining ordered, unordered, and partially ordered trees*, Knowl. and Inf. Syst., vol. 23, no. 5, pp.199-224, 2010.
[22] A. Ben-David, *Automatic Generation of Symbolic Multiattribute Ordinal knowledge-Based DSSs: methodology and Applications*, Decision Sciences, 1992, vol. 23, pp. 1357-137.
[23] S. Lievens, B. D. Baets, K. Cao-Van, *A probabilistic framework for the design of instance-based supervised ranking algorithms in an ordinal setting*, Annals of Operations Research, Vol. 163, No. 1, pp. 115-142, 2008
[24] B. Zhao, F. Wang, C. S. Zhang, *Block-quantized support vector ordinal regression*, IEEE Trans. Neural Netw., vol. 20, no. 5, pp. 882-890, 2009.
[25] B. Y. Sun, J. Y. Li, D. D.Wu, et al, *Kernel Discriminant Learning for Ordinal Regression*, IEEE Trans. Knowl. Data Eng., vol. 22, no. 6, pp. 906-910, 2010.
[26] I. Guyon and A. Elisseeff, *An introduction to variable and feature selection*, J. Mach. Learn. Res., vol. 3, pp. 1157-1182, 2003.
[27] H. Liu and L. Yu, *Toward integrating feature selection algorithms for classification and clustering*, IEEE Trans. Knowl. Data Eng., vol. 17, no. 4, pp. 491-502, Apr. 2005.
[28] R. Battiti, *Using mutual information for selecting features in supervised neural net learning*, IEEE Trans. Neural Netw., vol. 5, no. 4, pp. 537-550, Jul. 1994.
[29] H. Peng, F. Long, and C. Ding, *Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy*,IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226-1238, Aug. 2005.
[30] Q. H. Hu, D. R. Yu, and Z. X. Xie, *Information-preserving hybrid data reduction based on fuzzy-rough techniques*, Pattern Recognit. Lett., vol. 27, no. 5, pp. 414-423, Apr. 2006.
[31] Q. H. Hu, D. R. Yu, Z. X. Xie, J. F. Liu, *Fuzzy probabilistic approximation spaces and their information measures*, IEEE Trans. Fuzzy Syst., vol. 14, no. 2, pp. 191- 201, 2006.
[32] X. Hu and N. Cercone, *Learning in relational databases: A rough set approach*, Comput. Intell., vol. 11, no. 3, pp. 323-338, 1995.
[33] R. Jensen and Q. Shen, *Semantics-preserving dimensionality reduction: Rough and fuzzy-rough-based approaches*, IEEE Trans. Knowl. Data Eng., vol. 16, no. 12, pp. 1457-1471, Dec. 2004.
[34] Q. H. Hu, Z. X. Xie, D. R. Yu, *Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation*, Pattern Recog., vol. 40, 2007, pp. 3509 -3521.
[35] Q. H. Hu, D. R. Yu, J. F. Liu, C. X. Wu, *Neighborhood rough set based heterogeneous feature subset selection*, Inf. Sci., vol. 178, 2008, pp.3577-3594.
[36] R. Jensen and Q. Shen, *New Approaches to Fuzzy-Rough Feature Selection*, IEEE Trans. Fuzzy Syst., vol. 17, no. 4, pp. 824-838, 2009.
[37] M. Dash and H. Liu, *Consistency-based search in feature selection*, Artif. Intell., vol. 151, no. 1/2, pp. 155-176, Dec. 2003.
[38] Q. H. Hu, W. Pedrycz, D. R. Yu, J. Lang, *Selecting Discrete and Continuous Features Based on Neighborhood Decision Error Minimization*, IEEE Trans. Syst., Man, Cybern. B, Cybern.: Cybernetics, vol. 40, No. 1, 2010.
[39] N. Mac Parthalain, Q. Shen, R. Jensen, *A Distance Measure Approach to Exploring the Rough Set Boundary Region for Attribute Reduction*, IEEE Trans. Knowl. Data Eng., vol. 22, no. 3, pp. 306-317, 2010.
[40] M. R. Sikonja and I. Kononenko, *Theoretical and empirical analysis of ReliefF and RReliefF*, Mach. Learn., vol. 53, no. 1/2, pp. 23-69, Oct./Nov. 2003.
[41] Y. J. Sun, *Iterative RELIEF for feature weighting: Algorithms* theories, and applications," IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 6, pp. 1035-1051, Jun. 2007.
[42] Y. Sun, S. Todorovic, S. Goodison, *Local Learning Based Feature Selection for High-dimensional Data Analysis*, IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1610-1626, 2010.
[43] S. Nakariyakul, D. P. Casasent, *Adaptive branch and bound algorithm for selecting optimal features*, Pattern Recogn. Lett., vol. 28, no. 12, pp. 1415-1427, 2007.
[44] D. P. Muni, N. R. Pal, and J. Das, *Genetic programming for simultaneous feature selection and classifier design*, IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 36, no. 1, pp. 106-117, Feb. 2006.
[45] S. Nakariyakul, D. P. Casasent, *An improvement on floating search algorithms for feature subset selection*. Pattern Recogn., vol. 42,no. 9, pp. 1932-1940, 2009.
[46] T. Kamishima, S. Akaho, *Dimension Reduction for Supervised Ordering*, Sixth IEEE International Conference on Data Mining (ICDM'06), 2006, pp. 330-339.
[47] S. Baccianella, A. Esuli, F. Sebastiani, *Feature Selection for ordinal regression*, Proceeding SAC '10 Proceedings of the 2010 ACM Symposium on Applied Computing, pp. 1748-1754, 2010.
[48] W. H. Xu, X. Y. Zhang, J. M. Zhong, W. X. Zhang, *Attribute reduction in ordered information systems based on evidence theory*, Knowl. and Inf. Syst., vol. 25, no. 1, pp. 169-184, 2010.
[49] S. Greco, et al., *Variable consistency model of dominance-based rough sets approach*, W. Ziarko and Y. Yao (Eds.): RSCTC 2000, LNAI 2005, pp.170-181, 2001.
[50] J. Mingers. An empirical comparison of selection measures for decision-tree induction. Mach. Learn., 1989, vol. 3, no. 4, pp. 319-342.
[51] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, *The WEKA data mining software: an update*, SIGKDD Explorations, vol. 11, no. 1, pp. 10-18, 2009.
[52] R. Senge, E. Hullermeier, *Top-down induction of fuzzy pattern trees*, IEEE Trans. Fuzzy Syst., vol. 19, no. 2, pp. 241-252, 2011.
[53] E. Lughofer, S. Kindermann, *SparseFIS: data-driven learning of fuzzy systems With sparsity constraints*, IEEE Trans. Fuzzy Syst., vol. 18, no. 2, pp. 396-411 , 2010.
[54] Y. H. Qian, J. Y. Liang, W.-Z. Wu, C. Y. Dang. *Information granularity in fuzzy binary GrC model*, IEEE Trans. Fuzzy Syst., vol. 19, no. 2, pp. 253-264, 2011.
[55] Y. H. Qian, C. Y. Dang, J. Y. Liang, D. W. Tang. *Set-valued ordered information systems*, Inf. Sci., vol. 179, no. 16, pp. 2809-2832, 2009.
[56] G.-D. Wu, Z.-W. Zhu P.-H. Huang. *A TS-type maximizing-discriminability-based recurrent fuzzy network for classification problems*, vol. 19, no. 2, pp. 339-352, 2011.
[57] A. Frank, A. Asuncion, *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]*, Irvine, CA: University of California, School of Information and Computer Science, 2010.
[58] Q. H. Hu, D. R. Yu, W. Pedrycz, D. G. Chen, *Kernelized fuzzy rough sets and their applications*, IEEE Trans. Knowl. Data Eng., in press.
[59] Z. H. Deng, F.-L.Chung, S. T. Wang, *Robust relief-feature weighting, margin maximization, and fuzzy optimization*, IEEE Trans. Fuzzy Syst., vol.18, no. 4, pp. 726-744, 2010.

Qinghua Hu received B. E., M. E. and PhD degrees from Harbin Institute of Technology, Harbin, China in 1999, 2002 and 2008, respectively. Now he is an associate professor with Harbin Institute of Technology and a postdoctoral fellow with the Hong Kong Polytechnic University. His research interests are focused on intelligent modeling, data mining, knowledge discovery for classification and regression. He is a PC co-chair of RSCTC 2010 and severs as referee for a great number of journals and conferences. He has published more than 70 journal and conference papers in the areas of pattern recognition and fault diagnosis.

Weiwei Pan received her B.S from Tangshan Normal University and M.S. from Harbin Institute of Technology, Harbin, China in 2006 and 2009, respectively. Now she is a PhD candidate with Harbin Institute of Technology. Her interests are focused on monotonic classification, preference learning and their applications.



Lei Zhang received the B.S. degree in 1995 from the Shenyang Institute of Aeronautical Engineering, Shenyang, China, and the M.S. and Ph.D degrees in Electrical and Engineering from Northwestern Polytechnical University, Xi'an, China, respectively, in 1998 and 2001. From 2001 to 2002, he was a research associate with the Department of Computing, The Hong Kong Polytechnic University. From January 2003 to January 2006, he was a Postdoctoral Fellow in the Department of Electrical and Computer Engineering, McMaster University, Canada. Since January 2006, he has been an Assistant Professor in the Department of Computing, The Hong Kong Polytechnic University. His research interests include image and video processing, biometrics, pattern recognition, multisensor data fusion, machine learning and optimal estimation theory, etc.

David Zhang graduated in Computer Science from Peking University. He received his MSc in Computer Science in 1982 and his PhD in 1985 from the Harbin Institute of Technology (HIT). From 1986 to 1988 he was a Postdoctoral Fellow at Tsinghua University and then an Associate Professor at the Academia Sinica, Beijing. In 1994 he received his second PhD in Electrical and Computer Engineering from the University of Waterloo, Ontario, Canada. Currently, he is a Head, Department of Computing, and a Chair Professor at the Hong Kong Polytechnic University where he is the Founding Director of the Biometrics Technology Centre (UGC/CRC) supported by the Hong Kong SAR Government in 1998. He also serves as Visiting Chair Professor in Tsinghua University, and Adjunct Professor in Shanghai Jiao Tong University, Peking University, Harbin Institute of Technology, and the University of Waterloo. He is the Founder and Editor-in-Chief, International Journal of Image and Graphics (IJIG); Book Editor, Springer International Series on Biometrics (KISB); Organizer, the first International Conference on Biometrics Authentication (ICBA); Associate Editor of more than ten international journals including IEEE Transactions and Pattern Recognition; Technical Committee Chair of IEEE CIS and the author of more than 10 books and 200 journal papers. Professor Zhang is a Croucher Senior Research Fellow, Distinguished Speaker of the IEEE Computer Society, and a Fellow of both IEEE and IAPR.

Yanping Song She received her B.E., M. E. and PhD degrees from Harbin Institute of Technology, Harbin, China. Now she is a professor with School of Energy Science and Engineering, Harbin Institute of Technology. Her main interests are in modeling, simulation, and control of power systems.

Maozu Guo received his Bachelor and Master Degrees from Department of Computer Sciences, Harbin Engineering University in 1988 and 1991, respectively, and PhD from Department of Computer Sciences, Harbin Institute of Technology in 1997. Now he is director of the Natural Computation Division, Harbin Institute of Technology, Program Examining Expert of Information Science Division of NSFC; Senior Member of China Computer Federation (CCF), and Member of CCF Artificial Intelligence and Pattern Recognition Society; Member of Chinese Association for Artificial Intelligence(CAAI), and Standing Committee Member of Machine Learning Society of CAAI. Research Fields: machine learning and data mining; computational biology and bioinformatics; advanced computational models; image process and computer vision. Prof. GUO has implemented several projects from the Natural Science Foundation in China(NSFC), National 863Hi-tech Projects, the Science Fund for Distinguished Young Scholars of Heilongjiang Province, International Cooperative Project. He has won 1 Second Prize of the Province Science and Technology Progress, and 1 Third Prize of the Province Natural Science. He has published over 100 papers in journals and conferences.

Daren Yu received his M. E. and PhD degrees from Harbin Institute of Technology, Harbin, China, in 1988 and 1996, respectively. Since 1988, he has been working at the School of Energy Science and Engineering, Harbin Institute of Technology. His main research interests are in modeling, simulation, and control of power systems. He has published more than one hundred conference and journal papers on power control and fault diagnosis.