International Journal of Computational Intelligence Systems, Vol. 7, No. 3 (June 2014), 543-555

Feature selection for monotonic classification via maximizing monotonic dependency

Weiwei Pan, Qinghua Hu, Yanping Song, Daren Yu

Harbin Institute of Technology, Harbin 150001, China
E-mail: huqinghua@hit.edu.cn

Received 11 May 2012

Accepted 25 October 2013

Abstract

Monotonic classification is a special task in machine learning and pattern recognition. As to monotonic classification, it is assumed that both features and decision are ordinal and there is the monotonicity constraints between the features and decision. Little work has been focused on feature selection for this type of tasks although a number of feature selection algorithms have been introduced for nominal classification problems. However these techniques can not be applied to monotonic classification as they do not consider the monotonicity constraints. In this work, we present a technique to compute the quality of features for monotonic classification. Using gradient directing search method, this method trains a feature weight vector by maximizing the fuzzy monotonic dependency, which was defined in fuzzy preference rough sets. We conduct some experiments to compare the classification performances of the proposed method with some other techniques. The experimental results show the effectiveness of the proposed algorithm.

Keywords: Monotonic classification; feature selection; fuzzy monotonic dependency

1. Introduction

Monotonic classification is a kind of special ordinal classification, where the monotonicity constraints exist between the features and decision such that an object with better feature values should not be assigned to a worse decision. This kind of tasks occur in many real-world applications, such as bankruptcy risk [1], customer satisfaction analysis [2], house pricing [3], credit risk evaluation [4] and so on. Different assumptions of consistency are used in these two kinds of learning tasks. As to general classification problems, we think the samples with the same feature values should be assigned to the same class label; otherwise the samples are inconsistent. However, Under the monotonicity constraints, the object with better feature values should not be assigned to a worse class; otherwise, the decisions are not consistent. Compared to the general classification tasks, monotonic classification has not been fully investigated these years. However, it is attracting more and more attention in recent years [5,6,7,8].

It's well known that feature selection is one of important steps in machine learning and pattern recognition, which helps improve classification efficiency, avoid overfitting and enhance system generalization performance [9,10,11,12,13]. In pattern recognition and machine learning, most of learning tasks are described with a large number of features. However, some features are irrelevant or redundant [14,15]. A collection of feature selection algorithms have been designed for the general classification learning

Copyright: the authors

9,16,17,18,20,21,22,23,24,25,26,27,28,29,30.

As the feature evaluation measures used in general classification cannot reflect the monotonicity, they are not applicable to monotonic classification. We need to design new feature selection algorithms, which can describe the monotonicity consistency.

So far, some feature evaluation algorithms have been developed for monotonic classification [31,32,33,34]. Dominance-based rough set approach(DRSA) was firstly introduced by Greco, Matarazzo and Slowinski, where classical indiscernibility relations were replaced with dominance relations [1,35]. Recently, some fuzzy extension of dominance-based rough set approaches were proposed [36,37]. In 2010, Hu et al. introduced an algorithm to compute the fuzzy preference relations and constructed a fuzzy preference rough set model(FPRS) [37]. This model can deal with numerical or fuzzy features. In 2006, based on Kendall tau has been used to measure the correlation between two orders. Kamishima and Akaho introduced a method, called rank correlation dimension reduction, for a supervised ordering task [33]. In the same year, Xu, Zhang et al. extended the concepts of plausibility and belief reducts and applied them to attribute reduction in ordered information systems [34]. Qian et al. proposed attribute reductions of interval ordered decision tables and set-valued ordered decision tables [38,39]. In 2010, Baccianella, Esuli and Sebastiani designed four novel feature selection metrics for measuring the quality of features based on a scoring function for ordinal regression [6]. However, these scoring functions do not take the monotonicity constraints into consideration. By extending Shannon's entropy to monotonic classification, Hu et al. proposed new measures, called rank mutual information and fuzzy rank mutual information, to compute the relevance in monotonic classification [19,40]. In 2012, Hu et al. designed two feature evaluation functions for monotonic classification based on margin theory [46]. The experimental results verified the effectiveness and robustness of the proposed metrics.

The dependency, also called approximation quality, was defined as the ratio of consistent samples over the whole universe. It has been widely used to compute the quality of the features in general classification tasks [30]. The dependency function does not consider the monotonicity conditions. Then dependency was generalized to fuzzy monotonic dependency in the framework of fuzzy preference rough sets (FPRS). Fuzzy monotonic dependency can characterize the monotone consistency between features and decision, and it has been successfully used in attribute reduction [37,41]. It is clear that maximizing the fuzzy preference dependency between the features and decision means we get a monotonically consistent classification task. In this paper, we will design a feature selection algorithm via maximizing fuzzy preference dependency for monotonic classification. Different from the algorithm in [19,?], the gradient directing search algorithm is used to find the optimal feature weights. We compare the selected features with some other feature evaluation algorithms. It can be derived that the proposed algorithm is effective in measuring the monotone consistency.

The rest of the paper is organized as follows. In Section 2, the basic concepts of fuzzy preference rough sets (FPRS) are introduced . Then we design a feature weighting algorithm for monotonic classification through maximizing fuzzy monotonic dependency in Section 3. We show some experiments on the open datasets to demonstrate the effectiveness of the proposed algorithm in Section 4. Finally, conclusions are given in Section 5.

2. Fuzzy preference relation rough set model

In the rough set framework, objects are stored in a decision table, written as $S =< U, A, D >$, where $U = \{x_1, x_2, \cdots, x_n\}$ is a set of objects, $A = \{a_1, a_2, \cdots, a_J\}$ is a set of features, D is a ordinal set of decisions. Assume there are k classes, $D = \{d_1, d_2, \cdots, d_K\}$. Let x_{ij} represent the value of sample x_i on feature a_j. If the decision values are ordinal and discrete, without loss of generality, we assume that $d_1 \leqslant d_2 \leqslant \cdots \leqslant d_K$, d_k is better than d_{k-1}, we say $S =< U, A, D, f >$ is an ordinal classification task. Let $f : U \to D$ be the decision function, which assigns a decision value for each object in U.

Monotonic classification is a kind of special or-

Copyright: the authors

dinal classification, in which there exist the monotonicity constraint between the features and decision, that is for $\forall x_i, x_j \in U$, we have

$$x_i \geqslant x_j \Longrightarrow f(x_i) \geqslant f(x_j). \tag{1}$$

The monotonicity constraint can be interpreted as: for two samples x_i, x_j, if x_i with better features values than x_j should be classified into a decision class not worse than x_j's.

$\forall B \in A$, we say D is monotonically consistent with respect to B, if for $\forall x_i, x_j \in U$, we have $f(x_i) \geqslant f(x_j)$.

$\forall d_i \in D$, for the preference relations between the decision classes, we can derive the following nested preference decision structure.

$$d_i^{\leqslant} = \bigcup_{j=1}^{i} d_j, \tag{2}$$

where $d_i^{\leqslant}$ is a subset of samples whose decisions are equal to or less than d_i. If $i \leqslant j$, we have $d_i^{\leqslant} \subseteq d_j^{\leqslant}$ and $d_i^{\geqslant} \supseteq d_j^{\geqslant}$.

In fuzzy preference rough sets, for $\forall x_i, x_j \in U$, the fuzzy preference degree of x_i over x_j(the degree of x_i less than x_j) is computed as

$$r_{ij}^{<} = \frac{1}{1 + e^{k(f(x_i,a) - f(x_j,a))}}, \tag{3}$$

where k is a positive constant.

The logistic function $f(x) = 1/(1+\exp(-kx))$ is used to extract the preference relation [37]. Parameter k controls the preference degree defined by users. With different values of parameter k, the curves of the logistic function is shown in Fig.1. From Fig.1, we can see that when $k = 2, 5, 10, 20, 50$, the trends are the same . If k is too large, the logistic function becomes the membership function of classical set, the fuzzy preference rough set degenerates into dominance-based rough set. Some other transformation functions were also introduced in [42,43].

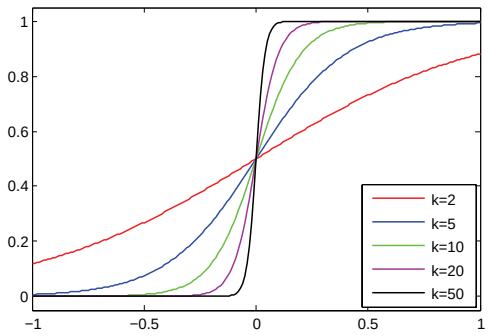

Fig. 1. The curves of the logistic function in interval [-1,1].

$r_{ij}^{<} = 0.5$, means there is no difference between x_i and x_j; $r_{ij}^{<} > 0.5$ denotes x_j is better than x_i; $r_{ij}^{<} < 0.5$ denotes x_i is greater than x_j.

For preference analysis, the fuzzy preference relation does not only reflect the preference structure between two samples, but also measures the degree of x_i less than x_j.

For $\forall x_i, x_j \in U$, we consider two features a_1 and a_2. Let $r_{ij}^{<}$ and $s_{ij}^{<}$ are two fuzzy preference relations derived with respect to feature a_1 and a_2, then the fuzzy preference relation of x_i over x_j based on feature subset (a_1, a_2) is defined as $\min(r_{ij}^{<}, s_{ij}^{<})$.

Given monotonic classification decision task $< U, A, D >$, $B \subseteq A$, $R^{<}$ is a fuzzy preference relation of sample x with respect to feature subset B. For $\forall d_i \subseteq D$, the lower approximation of preference decision classes $d_i^{\leqslant}$ is defined as

$$\underline{R^{<}} d_i^{\leqslant}(x) = \inf_{u \in U} \max(1 - R^{<}(u,x), d_i^{\leqslant}(u)). \tag{4}$$

The fuzzy lower approximation can be viewed as the certainty of x belongs to $d_i^{\leqslant}$, which reflects the decision consistency of the samples.

The approximation quality of D with respect to B is computed by the ratio of consistent samples over the the whole samples. It is also called dependency. Then we introduce a measure of dependency function for monotonic classification [37]. We call it monotonic dependency.

Given a monotonic classification decision table $< U, A, D >$, the decision D divides the universe into $d_1 \leqslant d_2 \leqslant \cdots \leqslant d_K$, $R^{<}$ is fuzzy preference relation

Copyright: the authors

induced with B, the monotonic dependency of decision D in terms of B is computed as

$$\gamma_B^{\preccurlyeq}(D^{\preccurlyeq}) = \frac{\sum_i \sum_{x \in d_i^{\preccurlyeq}} \underline{R^{\prec}} d_i^{\preccurlyeq}(x)}{\sum_i \left| d_i^{\preccurlyeq} \right|}, \tag{5}$$

where $|U|$ is the cardinality of U.

Monotonic dependency reflects the monotone consistency of the features with decision. It can be used for evaluating the quality of features in monotonic classification. If $\gamma_B^{\preccurlyeq}(D^{\preccurlyeq}) = 1$, we say D is monotonically consistent with B.

3. Feature selection via maximizing fuzzy preference dependency

3.1. Monotonic dependency

According to the definition of fuzzy preference rough sets [37], the fuzzy lower approximation operator of $d_i^{\preccurlyeq}$ is defined as $\underline{R^{\prec}} d_i^{\preccurlyeq}(x) = \inf_{u \in U} \max(1 - R^{\prec}(u,x), d_i^{\preccurlyeq}(u))$.

The lower approximation of preference decision classes can be computed with

If $u \in d_i^{\preccurlyeq}$, we have $max(1 - R^{\prec}(u,x), d_i^{\preccurlyeq}(u)) = 1$.

Similarly, if $u \notin d_i^{\preccurlyeq}$, that is $d_i^{\preccurlyeq}(u) = 0$, hence $max(1 - R^{\prec}(u,x), d_i^{\preccurlyeq}(u)) = 1 - R^{\prec}(u,x)$. Since $R^{\prec}(u,x) < 1$, so $\underline{R^{\prec}} d_i^{\preccurlyeq}(x) = \inf_{u \notin d_i^{\preccurlyeq}} 1 - R^{\prec}(u,x)$.

Thus, we can rewritten the fuzzy lower approximation of $d_i^{\preccurlyeq}$ as $\underline{R^{\prec}} d_i^{\preccurlyeq}(x) = \inf_{u \notin d_i^{\preccurlyeq}} 1 - R^{\prec}(u,x)$.

This expression suggests that the membership of sample x to the fuzzy preference lower approximation of $d_i^{\preccurlyeq}$ is computed with the sample whose decision label larger than d_i and has the greatest preference degree over x.

In fact, we can compute the sample and denote it as y.

Let $w = (w_1, w_2, \cdots w_J)'$ be a nonnegative feature weight vector of the features, J is the feature dimension. N is the number of samples in U. $w \geqslant 0$. This constraint guarantees that the features monotonically increase with the decision.

Then the fuzzy preference lower approximation with respect to w can be computed as:

$$\underline{R^{\prec}} d_i^{\preccurlyeq}(x) = \frac{1}{1 + e^{kw^T(x-y)}}, \tag{6}$$

where $w^T(x-y) = \sum_{j=1}^{J} w_j (x-y)^j$.

Then we have the fuzzy preference dependency of D on B in the following form:

$$\gamma_B^{\preccurlyeq}(D^{\preccurlyeq}) = \frac{\sum_{n=1}^{N} \frac{1}{1+e^{kw^T(x_n - y_n)}}}{\sum_i \left| d_i^{\preccurlyeq} \right|}. \tag{7}$$

3.2. Maximizing monotonic dependency by gradient descent

Monotonic dependency reflects the monotone consistency between the features and decision. If we maximize the value of monotonic dependency, we can obtain a more monotonically consistent classification task. Based on the definition of monotonic dependency, maximizing the value of monotonic dependency equals to maximizing the sum of fuzzy lower approximation of the universe.

The evaluation function is written as

$$e(w) = \sum_{n=1}^{N} \underline{R^{\prec}} d_i^{\preccurlyeq}(x_n) = \sum_{n=1}^{N} \frac{1}{1 + e^{kw^T(x_n - y_n)}}. \tag{8}$$

We obtain the following optimization objective function.

$$\max_w e(w) =$$

$$\sum_{n=1}^{N} \frac{1}{1 + \left(e^{kw_1(x_n - y_n)^1} \times e^{kw_2(x_n - y_n)^2} \cdots \times e^{kw_I(x_n - y_n)^I}\right)}$$

.

This formulation is equal to

$$\min_w \sum_{n=1}^{N} 1 + \left(e^{kw_1(x_n - y_n)^1} \times e^{kw_2(x_n - y_n)^2} \cdots \times e^{kw_I(x_n - y_n)^I}\right)$$

.

The optimization objective function can be written as

$$\min_w \sum_{n=1}^{N} \left(e^{kw_1(x_n - y_n)^1} \times e^{kw_2(x_n - y_n)^2} \cdots \times e^{kw_I(x_n - y_n)^I}\right) \tag{9}$$

$$= \min_w \sum_{n=1}^{N} e^{kw^T(x_n - y_n)}. \tag{10}$$

Copyright: the authors

Because of the nonnegative constraint on feature weight vector w, we assume that $w_j = v_j^2, 1 \leqslant j \leqslant J$. Now, for the fixed x_n, the above expression is smoothing almost everywhere. Thus we can solve this problem by gradient descent and reach an optimal feature subsets. The update rule is formulated as

$$\nabla e(w))_i =$$

$$v - \eta(k \sum_{n=1}^{N} \exp(k \sum_{j} v_j^2 (x_n - y_n)^j (x_n - y_n))) \otimes v. \quad (11)$$

In each iteration, with the last weight vector w, a new feature weight vector is computed through gradient descent. It operates iteratively until the stopping condition is satisfied. Then we get an optimum feature weight vector. The pseudocode of the proposed algorithm is described as follows.

initialize the weight vector: $w =< 1, 1, \cdots, 1 >$, t=1;
stopping criterion θ;
parameter k in the logsig function;
for each $x_n \in U$, assume its decision is d_i,
find the sample which does not belong the the class $d_i^{\leqslant}$,
and has the least feature value corresponding to x_n,
denoted by y_n;
compute v through gradient descent, the equation is
$v \leftarrow v - \eta(k \sum_{n=1}^{N} \sum_{i} \exp(k \sum_{j} v_j^2 (x_n - x_i)^j (x_n - y_n))) \otimes v$
$w^t \leftarrow v^2, 1 \leqslant j \leqslant J$
if $\left\| w^{(t)} - w^{(t-1)} \right\| < \theta$
return w
else $t \leftarrow t + 1$
go to line 3
end
$w = w - \min(w) / (\max(w) - \min(w))$
end

The computational complexity of the proposed algorithm is $O(N^2 J)$, where N is the number of samples in U and J is the feature dimensionality.

Running this algorithm, we can obtain the weights of different features. Then we rank these features in a descending order and we add the features into the selected pool one by one. We observe the variation of classification performance when new features are added. Based on the definition of monotonic dependency, a forward greedy search algorithm was constructed for feature selection in [37]. This algorithm converges to a local minimum and the time complexity is $O((N \log N + KN)J^2)$, which is worse than the proposed algorithm.

4. Experiments

In this section, we conduct some experiments to show the effectiveness of the proposed algorithm. We collect eleven monotonic classfication tasks from UCI machine learning repository and WEKA homepage. The detailed information information of these datasets are listed in Table 1.

Table 1: Data description.

Data set	Instances	Features	Classes
Ailerons	13750	40	3
Australian credit	690	14	2
Autompg	392	7	3
Cardiotocography	2126	21	3
German credit	1000	20	2
Japan credit	690	15	2
Housing	506	13	4
Pasture	36	22	3
Triazines	186	61	3
Wdbc	569	30	2
Wine quality-red	1599	11	6

We run our algorithm(MFW-MMD) along with two other metrics, fuzzy preference rough set(FPRS) [37] and fuzzy rank mutual information(FRMI) [19,40,47]. The proposed algorithm(MFW-MMD) is designed specially for monotonic classification by defining an optimization objective function based on fuzzy preference dependency. This function can characterize the monotone consistency between the features and decision. Extended from Shannon information theory by integrating fuzzy preference relations into information entropy, fuzzy rank mutual information(FRMI) also characterizes the monotone relevance between the features and decision and has been proven its effectiveness in selecting monotonic features. All these three feature weighting methods can be used for monotonic classification.

First, we study the convergence of the proposed algorithm on some datasets. We compute the difference $\theta = \left\| w^{(t)} - w^{(t-1)} \right\|$ between the weight vec-

Copyright: the authors

tor when new iteration is completed. The results are shown in Fig.2. We set the stopping criterion $\theta = 0.01$.

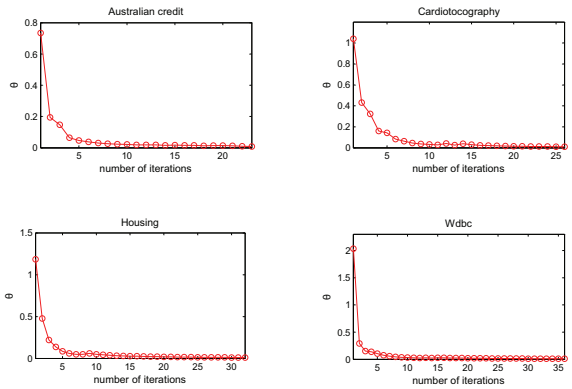

Fig. 2. The convergence analysis of the proposed algorithm.

Fig.3 shows there is not significant difference between the weights computed in the last iteration and the second last iteration.

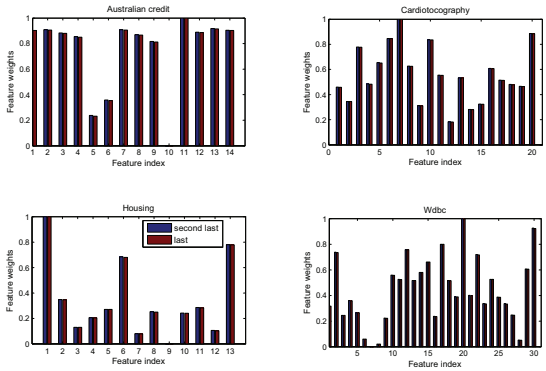

Fig. 3. The last and second last feature weights.

Fig.2 and Fig.3 show that the difference of weights converges after several iterations. So the proposed algorithm can get stable weight vectors.

Now we compare the classification accuracies obtained with different algorithms. We first train the weight vector, and then rank the features in a descending order. The classification accuracy is calculated by adding the top ranked features into classification algorithms one by one.

For each dataset, we compared the classification accuracies of the proposed algorithm against the above two other evaluation metrics utilizing four classifiers: nearest neighbors(KNN), classification and regression trees(CART), Rank Tree [44] and Ordinal Stochastic Dominance Learner(OSDL) [45]. The first two classifiers do not consider the monotonicity constraints, whereas Rank Tree and OSDL are two monotonic classifiers. For each dataset and algorithm, classification performances are estimated with 10-fold cross validation. Then we get the optimum feature subsets with respect to the maximum value of classification accuracies.

In this paper, we set the parameter of the preference metric $k = 10$. The classification results are shown in Tables 2 to 5, where the best classification performance are marked with bold and the numbers of the selected features are given in bracket, the last column gives the classification accuracy with the raw data. From the results, we can conclude that the classification performance of datasets are improved compared with the raw data after feature selection. It demonstrates that our algorithm is effective for monotonic classification. In addition, for each dataset, compared to FRMI and FPRS, our new proposed method(MFW-MMD) also obtains the best performances with respect to the four classification algorithms, respectively. These results show that our algorithm(MFW-MMD) can characterize the monotonicity structure between the features and decision, and it improves the classification performances.

In order to compare these algorithms, We give a statistical test of the performances between MFW-MMD, FPRS and FRMI based on the Friedman test [48]. The performance of two classifiers is significantly different if the corresponding average ranks differ by at least the critical difference

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{12}$$

In this paper, $q_{0.05} = 2.343$, thus $CD = 1.00$. We compute Table 2, the average ranks for MFW-MMD, FPRS and FRMI are 1.18, 2.23, 2.50, respectively. The differences of the average rank between MFW-MMD and the other methods are $2.23 - 1.18 = 1.05 > 1.00$, $2.50 - 1.18 = 1.32 > 1.00$. Thus MFW-MMD performs significantly better than FPRS and FRMI. We also check the re-

Copyright: the authors

Table 2: Comparison of three evaluation functions with KNN(1-nearest neighbor).

Data set	MFW-MMD	FPRS	FRMI	Raw data
Ailerons	**58.74±7.70(17)**	57.16±8.71(19)	57.26±10.13(38)	56.59±5.77
Australian credit	**87.26±4.76(4)**	84.79±3.07(8)	84.35±4.47(9)	80.98±5.06
Autompg	**79.87±8.64(2)**	79.48±7.46(3)	77.41±7.36(4)	54.99±14.32
Cardiotocography	**86.41±10.12(5)**	84.71±8.99(6)	85.13±8.62(10)	83.84±5.24
German credit	**72.00±3.30(15)**	71.70±3.43(13)	70.70±2.71(17)	69.70±3.80
Housing	**67.48±7.26(4)**	66.75±10.33(10)	63.65±9.18(12)	63.50±8.87
Japan credit	83.76±16.55(15)	**85.08±12.83(9)**	85.06±14.30(9)	83.76±15.27
Pasture	**85.56±16.60(11)**	82.22±19.03(14)	82.22±19.03(14)	67.78±20.59
Triazines	**57.97±17.76(33)**	57.92±8.02(30)	52.50±9.30(40)	50.56±6.65
Wdbc	**97.37±2.07(26)**	97.20±2.36(30)	97.20±2.36(29)	97.20±2.36
Wine quality-red	**51.44±9.74(8)**	51.05±7.16(9)	51.05±7.16(9)	47.22±4.34

Table 3: Comparison of three evaluation functions with CART.

Data set	MFW-MMD	FPRS	FRMI	Raw data
Ailerons	**65.49±5.01**(14)	63.37±8.40(38)	63.31±9.43(38)	57.58±6.31
Australian credit	**86.39±4.99(5)**	85.52±5.20(2)	85.52±5.20(1)	82.60±4.45
Autompg	**87.41±6.69(3)**	87.41±6.89(4)	86.63±7.97(3)	81.25±7.09
Cardiotocography	**86.73±10.04(5)**	86.40±8.89(6)	85.73±10.98(21)	85.59±10.78
German credit	**71.30±3.83(14)**	70.10±5.07(19)	71.10±4.72(18)	69.90±3.54
Housing	**64.09±7.99(7)**	64.04±7.35(11)	63.66±10.06(11)	62.11±8.86
Japan credit	82.02±14.29(7)	83.30±16.80(2)	**85.48±18.51(1)**	82.73±14.86
Pasture	**83.33±17.57(4)**	77.78±15.71(2)	81.11±16.60(4)	64.44±17.21
Triazines	**58.89±7.29(35)**	54.86±6.12(31)	54.86±9.61(24)	48.06±11.11
Wdbc	93.49±5.44(17)	**94.21±3.42(5)**	92.79±4.17(15)	90.50±4.55
Wine quality-red	**55.51±10.97(6)**	52.05±6.51(7)	53.98±14.07(7)	51.53±2.89

Table 4: Comparison of three evaluation functions with Rank Tree.

Data set	MFW-MMD	FPRS	FRMI	Raw data
Ailerons	**62.91±7.00(29)**	59.10±10.57(11)	61.34±7.69(14)	62.82±3.85
Australian credit	**70.60±4.92(9)**	55.50±0.86(1)	55.50±0.86(1)	55.50±0.86
Autompg	**81.22±6.98(4)**	80.97±6.99(4)	80.45±7.01(7)	80.45±7.01
Cardiotocography	**83.51±10.60(9)**	81.54±9.74(21)	83.51±10.60(9)	81.54±9.74
Housing	**67.80±8.24(10)**	65.90±8.69(13)	65.32±10.25(13)	65.08±8.28
Japan credit	**71.73±5.51(4)**	67.53±4.40(1)	55.08±4.82(1)	55.08±4.82
Pasture	**81.11±16.60(4)**	73.33±26.29(2)	75.56±21.47(4)	71.11±22.95
Triazines	50.00±6.42(12)	48.75±9.18(52)	**55.56±4.54(49)**	47.08±9.32
Wdbc	**93.68±3.00(16)**	93.33±3.29(2)	93.48±3.81(3)	92.62±3.10
Wine quality-red	**56.95±9.10(9)**	54.57±9.13(9)	56.14±9.87(8)	55.10±5.59

Copyright: the authors

Table 5: Comparison of three evaluation functions with OSDL

Data set	MFW-MMD	FPRS	FRMI	Raw data
Ailerons	**67.83±4.38(25)**	67.07±4.94(31)	67.83±5.30(6)	66.96±6.16
Australian credit	**85.80±1.42(5)**	85.51±1.45(1)	82.03± 1.80(8)	79.28±2.07
Autompg	**90.31±5.77(5)**	86.99 ± 5.79(7)	88.52 ± 5.79(6)	88.52±5.79
Cardiotocography	**87.96±1.70(4)**	85.09± 1.68(4)	83.35±2.17(4)	73.52±3.01
German credit	**71.60±2.84(3)**	71.00±2.90(7)	70.00±3.00(1)	49.10±5.09
Housing	**54.35±6.76(10)**	53.95± 6.88(12)	52.77± 7.00(12)	42.09±7.95
Japan credit	65.07±3.50(13)	65.07±3.50(15)	**65.36± 3.46(15)**	55.94±4.4
Pasture	**91.67±0.83(8)**	80.56±1.94(6)	77.78± 2.22(3)	75.00 ±3.06
Triazines	**79.03 ±2.37(48)**	48.92± 6.99(22)	47.31 ±6.45(15)	42.47±1.08
Wdbc	**71.00 ±2.90(16)**	62.74±3.73(16)	62.74±3.73(22)	62.21 ±3.78
Wine quality-red	**58.84±3.45(5)**	57.91±3.58(6)	58.47±3.69(4)	54.97±2.77

sults in Table 3 to Table 5. In Table 3, the differences of the average rank between MFW-MMD and FPRS are $2.23 - 1.32 = 0.91 < 1.00$, the differences of the average rank between MFW-MMD and FRMI are $2.45 - 1.32 = 1.13 > 1.00$, thus MFW-MMD performs significantly better than FRMI and the difference of MFW-MMD and FPRS is not significant based on CART. For Rank Tree, the differences of the average rank between MFW-MMD and FPRS are $2.65 - 1.15 = 1.50 > 1.00$, the differences of the average rank between MFW-MMD and FRMI are $2.20 - 1.15 = 1.05 > 1.00$, the results indicate MFW-MMD behaves much better than FPRS and FRMI. Then for OSDL, the differences of the average rank between MFW-MMD and FPRS are $2.36 - 1.18 = 1.18 > 1.00$, the differences of the average rank between MFW-MMD and FRMI are $2.45 - 1.18 = 1.27 > 1.00$, MFW-MMD behaves much better than FPRS and FRMI.

The classification performance varying with number of the selected features are given in Figs. 4 to 7. From these figures, we obtain the performances of almost all of the classification tasks are improved firstly when we add new features, however, some of them then drop or keep invariable if more features are selected. Thus a better classification performance can be reached if a proper feature subset is selected.

In order to show the effectiveness of the algorithm, we finally give the class distribution over the features with weights. Take wdbc and housing as examples. The probability density of the best features are given in Fig.8. We plot the probability density distribution using Parzen window estimation. Four features are considered, which get the largest feature weight learned with our algorithm. Fig.9 shows the scatter plot with the first larger weights features of housing data: feature 1 and feature 13, and the probability density function is plotted based on feature 13. Although there are some inconsistent samples in data, the probability density function of each class satisfies the monotonicity constraints relative to the given feature.

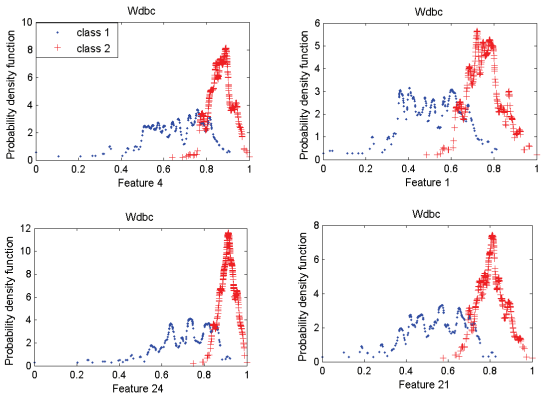

Fig. 8. The probability density function of four features with the largest feature weight on the wdbc data.

Copyright: the authors

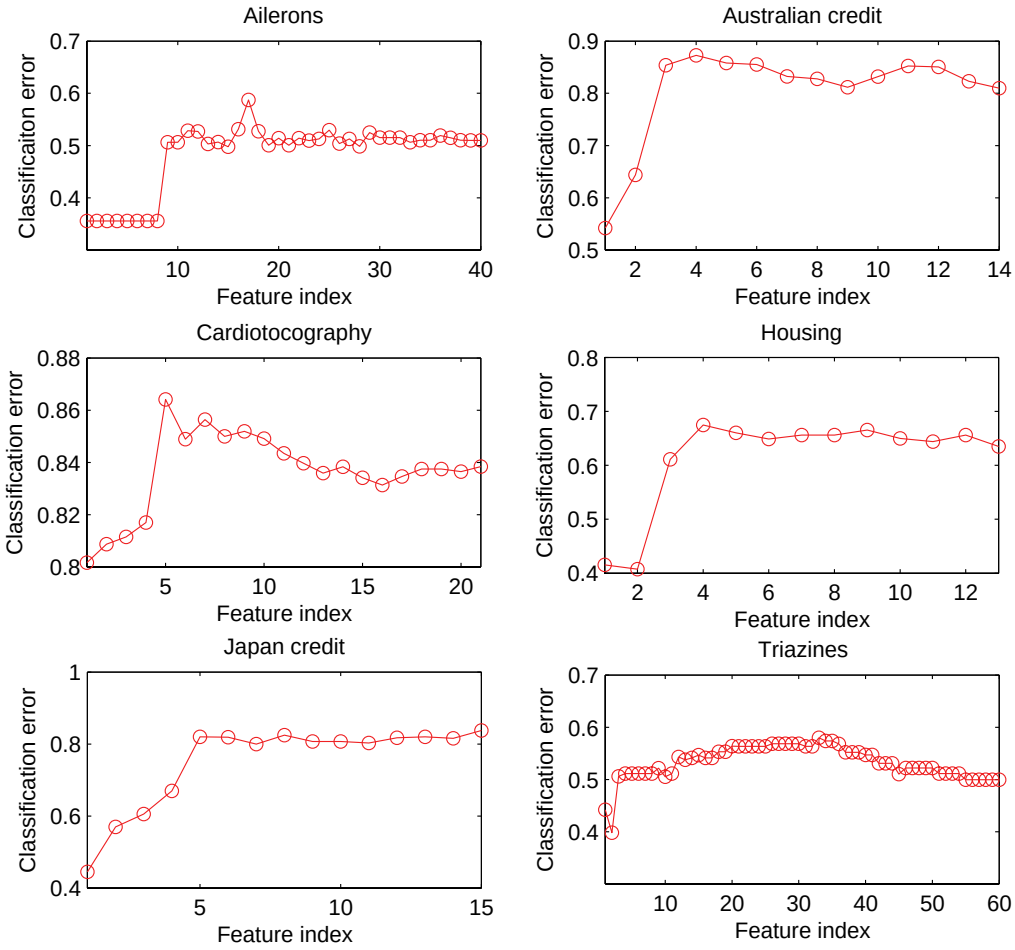

Figure 4: Classification accuracy on six datasets base on KNN.

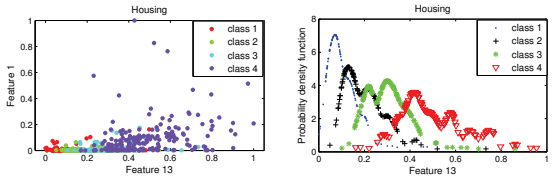

Fig. 9. The experiments on the housing data.

5. Conclusions

Monotonic classification is a kind of special classification tasks in decision analysis. In this work, we design a feature selection algorithm for monotonic classification. Fuzzy preference dependency defined in fuzzy preference rough set reflects the monotone consistency of each feature, and has been used to evaluate the quality of the features. In this paper, we maximize the fuzzy preference dependency by gradient descent. Then a feature weight vector is learned.

Copyright: the authors

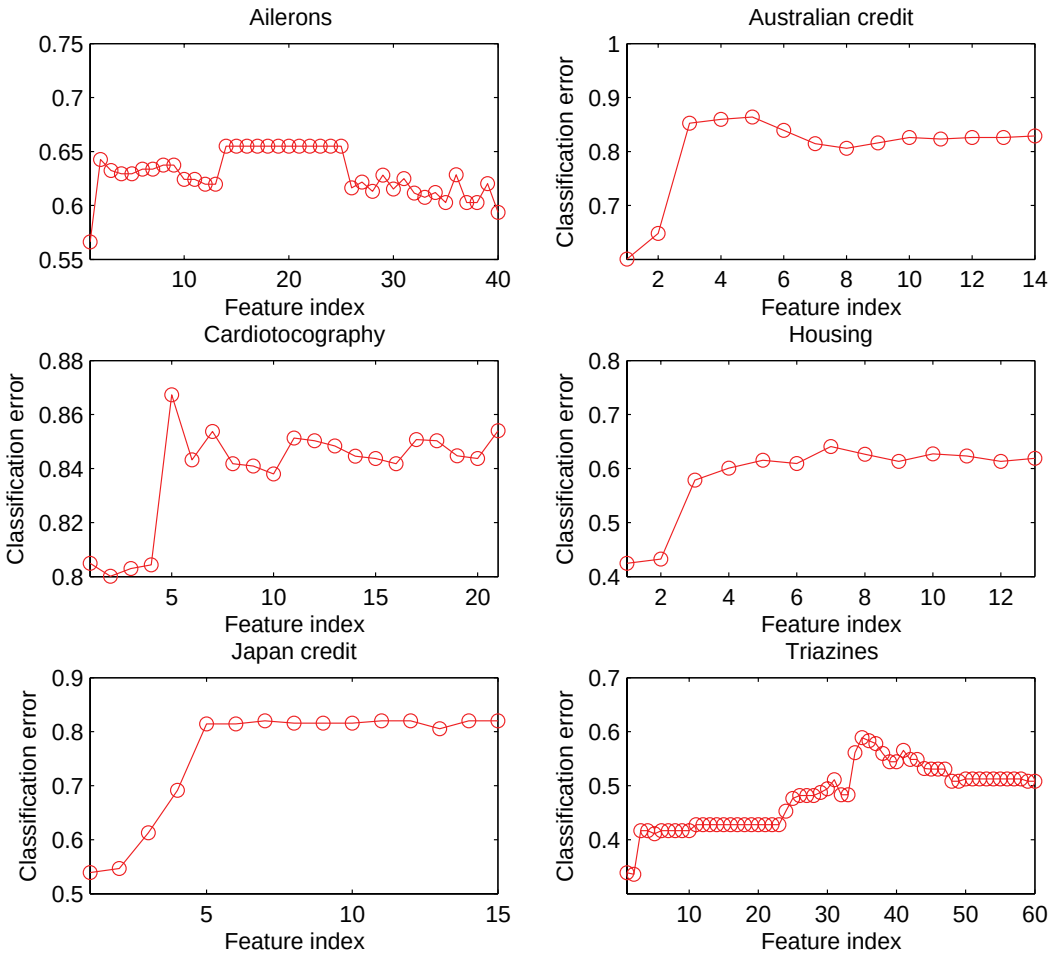

Figure 5: Classification accuracy on six datasets base on CART.

We conduct some experiments on real-world datasets and give comparisonal analysis on some related algorithms. According to the experimental results, we derived that our algorithm outperforms fuzzy preference rough set(FPRS) and fuzzy rank mutual information(FRMI) based forward greedy search.

Acknowledgments

This work is partially supported by National Natural Science Foundation of China under Grants 61222210 and 10978011.

References

1. S. Greco, B. Matarazzo and R. Slowinski, "Rough approximation of a preference relation by dominance relations," European Journal of Operational Research, 25, 327-341(2005).
2. S. Greco, B. Matarazzo and R. Slowinski, "Customer satisfaction analysis based on rough set approach," Zeitschrift für Betriebswirtschaft, 77, 325-339(2007).
3. R. Potharst and A. Feelders, "Classification trees for problems with monotonicity constraints," ACM SIGKDD Explorations Newsletter, 4, 1-10(2002).

Copyright: the authors

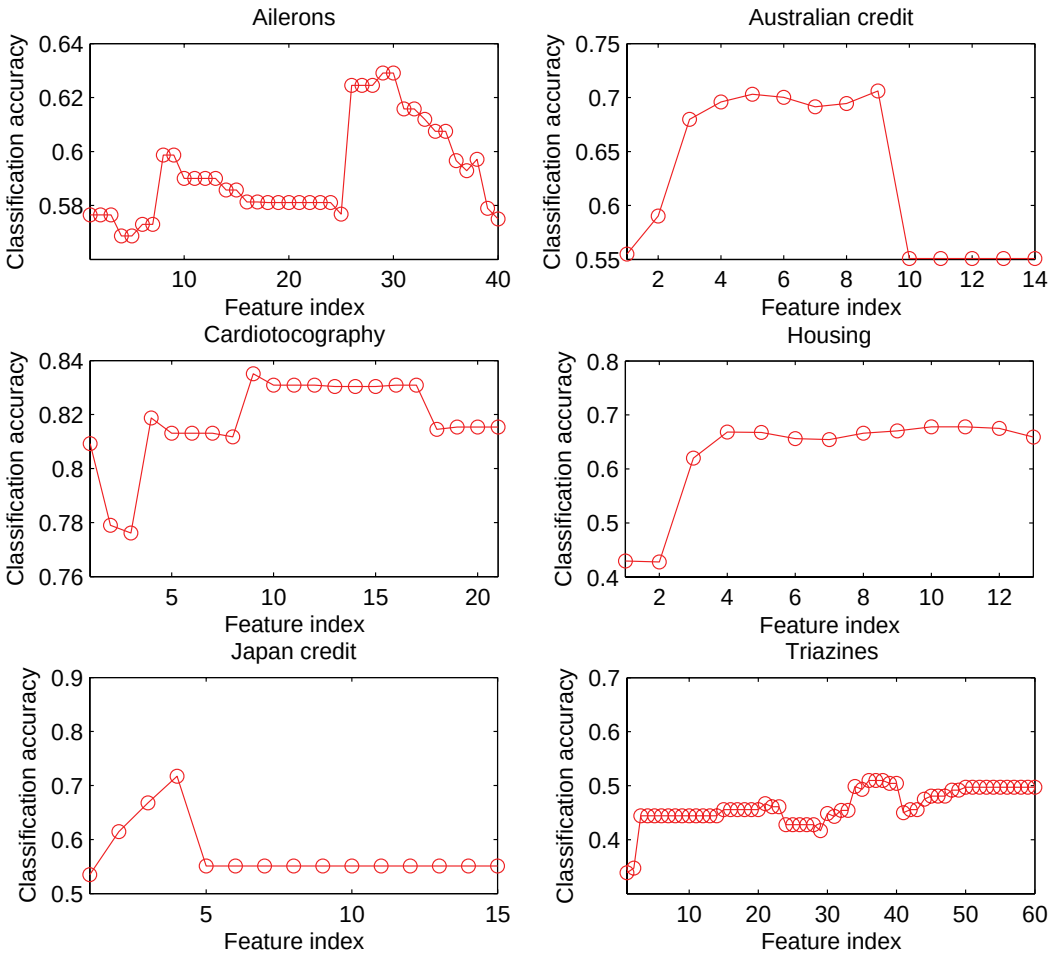

Figure 6: Classification accuracy on six datasets base on Rank Tree.

4. M. Doumpos and F. Pasiouras, "Developing and testing models for replicating credit ratings: A multicriteria approach," Computational Economics, 25, 327-341(2005).
5. A. Ben-David, L. Sterling and T.D. Tran, "Adding monotonicity to learning algorithms may impair their accuracy," Expert Systems with Applications, 36, 6627-6634(2009).
6. J. Błaszczyński, R. Słowiński and J. Stefanowski, "Ordinal classification with monotonicity constraints by variable consistency bagging," Rough Sets and Current Trends in Computing, 392-401(2010).
7. W. Duivesteijn and A. Feelders, "Nearest neighbour classification with monotonicity constraints," Machine Learning and Knowledge Discovery in Databases, 301-316(2008).
8. R. Sousa, I. Yevseyeva, JFP.da Costa, et al, "Multicriteria models for learning ordinal data: a literature review," Artificial Intelligence, Evolutionary Computing and Metaheuristics, Springer Berlin Heidelberg, 109-138(2013).
9. R. Gilad-Bachrach, A. Navot and N. Tishby, "Margin based feature selection-theory and algorithms," In Proceedings of the 21th International Conference on Machine Learning, 43-50(2004).
10. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," The Journal of Machine Learning Research, 3, 1157-1182(2003).

Copyright: the authors

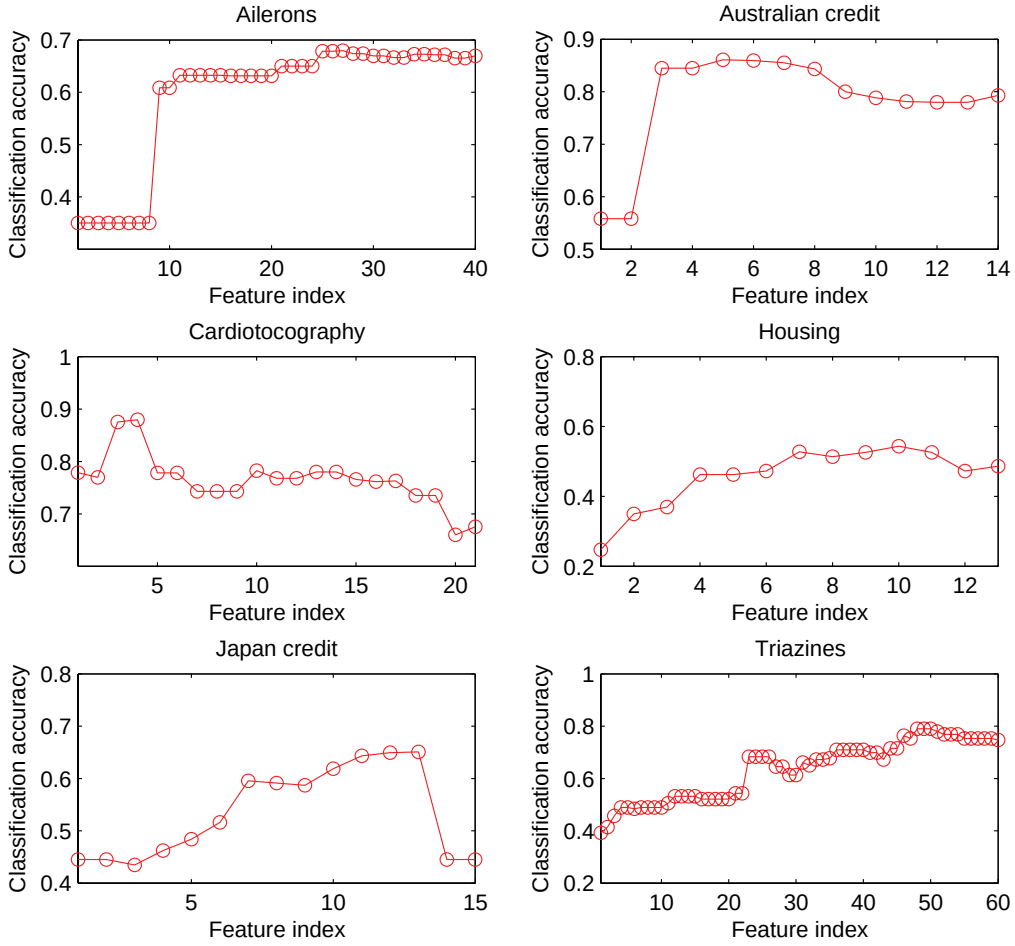

Figure 7: Classification accuracy on six datasets base on OSDL.

11. A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," IEEE Transactions on Pattern Analysis and Machine Intelligence, 19, 153-158(1997).
12. N. Kwak and C.H. Choi, "Input feature selection for classification problems," IEEE Transactions on Neural Networks, 13, 143-159(2002). Kalousis
13. D.P. Muni, N.R. Pal and J. Das, "Genetic programming for simultaneous feature selection and classifier design," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 36, 106-117(2006).
14. Q.H. Hu, Z.X. Xie and D.R. Yu, "Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation," Pattern Recognition, 40, 3509-3521(2007).
15. Q.H. Hu, D.R. Yu, J.F. Liu and C.X. Wu, "Neighborhood rough set based heterogeneous feature subset selection," Information sciences, 178, 3577-3594(2008).
16. M. Hilario and A. Kalousis, "Approaches to dimensionality reduction in proteomic biomarker studies," Briefings in bioinformatics, 9, 102-108(2008).
17. R. Kohavi and G.H. John, "Wrappers for feature subset selection," Artificial intelligence, 97, 273-324(1997).
18. P. Pudil and J. Hovovicova, "Novel methods for subset selection with respect to problem knowledge," IEEE Transactions on Intelligent Systems and their Applications, 13, 66-74(1998).

Copyright: the authors

19. Q.H. Hu, M.Z. Guo, D.R. Yu and J.F. Liu, "Information entropy for ordinal classification," SCIENCE CHINA Information Sciences, 53, 1188-1200(2010).
20. 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, 27, 1226-1238(2005).
21. D. Slezak, "Approximate entropy reducts," Fundamenta Informaticae, 53, 365-390(2002).
22. G.Y. Wang, J. Zhao, J. An and Y. Wu, "A comparative study of algebra viewpoint and information viewpoint in attribute reduction," Fundamenta Informaticae, 68, 289-301(2005).
23. B. Chen, H. Liu, J. Chai and Z. Bao, "Large margin feature weighting method via linear programming," IEEE Transactions on Knowledge and Data Engineering, 21, 1475-1488(2009).
24. Y. Huang, P.J. McCullagh and N.D. Black, "An optimization of ReliefF for classification in large datasets," Data and Knowledge Engineering, 68, 1348-1356(2009).
25. Y. Sun, "Iterative RELIEF for feature weighting: algorithms, theories, and applications," IEEE Transactions on Pattern Analysis and Machine Intelligence, 29, 1035-1051(2007).
26. Y. Sun, S. Todorovic and S. Goodison, "Local-learning-based feature selection for high-dimensional data analysis," 32, 1610-1626(2010).
27. Q.H. Hu, S. An and D.R. Yu, "Soft fuzzy rough sets for robust feature evaluation and selection," Information Sciences, 180, 4384-4400(2010).
28. R. Jensen and Q. Shen, "Fuzzy-rough sets assisted attribute selection," IEEE Transactions on Fuzzy Systems, 15, 73-89(2007).
29. M. Modrzejewski, "Feature selection using rough sets theory," Machine Learning: ECML-93, 213-226(1993).
30. R.W. Swiniarski and A. Skowron, "Rough set methods in feature selection and recognition," Pattern recognition letters, 24, 833-849(2003).
31. S. Baccianella, A. Esuli and F. Sebastiani, "Feature selection for ordinal regression," In Proceedings of the 2010 ACM Symposium on Applied Computing, 1748–1754(2010).
32. S. Greco, B. Matarazzo, R. Slowinski and J. Stefanowski, "Variable Consistency Model of Dominance-Based Rough Sets Approach," Rough Sets and Current Trends in Computing, 170-181(2001).
33. T.Kamishima and S. Akaho, "Dimension reduction for supervised ordering," ICDM'06. Sixth International Conference on Data Mining, 330-339(2006).
34. W. Xu, X. Zhang, J. Zhong and W. Zhang, "Attribute reduction in ordered information systems based on evidence theory," Knowledge and Information Systems, 25, 169-184(2010).
35. S. Greco, B. Matarazzo and R. Slowinski, "Rough approximation by dominance relations," International journal of intelligent systems, 17, 153-171(2002).
36. S. Greco, B. Matarazzo and R. Slowinski, "Fuzzy extension of the rough set approach to multicriteria and multiattribute sorting," Studies in Fuzziness and Soft Computing, 51, 131-152(2000).
37. Q.H. Hu, D.R. Yu and M.Z. Guo, "Fuzzy preference based rough sets," Information Sciences, 180, 2003-2022(2010).
38. Y. Qian, C. Dang, J. Liang and D. Tang, "Set-valued ordered information systems," Information Sciences, 179, 2809-2832(2009).
39. Y. Qian, J. Liang and C. Dang, "Interval ordered information systems," Computers and Mathematics with Applications, 56, 1994-2009(2008).
40. Q.H. Hu, W.W. Pan, L. Zhang, D. Zhang, Y.P. Song, M.Z. Guo and D.R. Yu," Feature selection for monotonic classification," IEEE Transactions on Fuzzy Systems, 20, 69-81(2012).
41. Z. Xu, "Consistency of interval fuzzy preference relations in group decision making," Applied Soft Computing, 11, 3898-3909(2011).
42. T.F. Fan, C.J. Liau and D.R. Liu, "Dominance-based fuzzy rough set analysis of uncertain and possibilistic data tables," International Journal of Approximate Reasoning, 52, 1282-1297(2011).
43. S. Greco, B. Matarazzo and R. Słowinski, "Fuzzy Set Extensions of the Dominance-Based Rough Set Approach," Studies in Fuzziness and Soft Computing, 220, 239-261(2008).
44. F. Xia, W. Zhang, F. Li and Y. Yang, "Ranking with decision tree," Knowledge and Information Systems, 17, 381-395(2008).
45. S. Lievens, B. De Baets and K. Cao-Van, "A probabilistic framework for the design of instance-based supervised ranking algorithms in an ordinal setting," Annals of Operations Research, 163, 115-142(2008).
46. Q. Hu, W. Pan, Y. Song, et al, "Large-margin feature selection for monotonic classification," Knowledge-Based Systems, 31, 8-18(2012).
47. Q. Hu, X. Che, L. Zhang L, et al, "Rank entropy based decision trees for monotonic classification," IEEE Transactions on Knowledge and Data Engineering, 24, 2052-2064(2012).
48. J. Demsar J, "Statistical comparisons of classifiers over multiple data sets," The Journal of Machine Learning Research, 7, 1-30(2006).

Copyright: the authors