SCIENCE CHINA
Information Sciences

• RESEARCH PAPERS •

June 2010 Vol. 53 No. 6: 1188–1200
doi: 10.1007/s11432-010-3117-7

Information entropy for ordinal classification

HU QingHua*, GUO MaoZu, YU DaRen & LIU JinFu

Harbin Institute of Technology, Harbin 150001*, China*

Received April 20, 2009; accepted June 30, 2009

Abstract Ordinal classification plays an important role in various decision making tasks. However, little attention is paid to this type of learning tasks compared with general classification learning. Shannon information entropy and the derived measure of mutual information play a fundamental role in a number of learning algorithms including feature evaluation, selection and decision tree construction. These measures are not applicable to ordinal classification for they cannot characterize the consistency of monotonicity in ordinal classification. In this paper, we generalize Shannon's entropy to crisp ordinal classification and fuzzy ordinal classification, and show the information measures of ranking mutual information and fuzzy ranking mutual information. We discuss the properties of these measures and show that the proposed ranking mutual information and fuzzy ranking mutual information are the indexes of consistency of monotonicity in ordinal classification. In addition, the proposed indexes are used to evaluate the monotonicity degree between features and decision in the context of ordinal classification.

Keywords ordinal classification, information entropy, ranking entropy, ranking mutual information

Citation Hu Q H, Guo M Z, Yu D R, et al. Information entropy for ordinal classification. Sci China Inf Sci, 2010, 53: 1188–1200, doi: 10.1007/s11432-010-3117-7

1 Introduction

Ordinal classification (also called ranking or sorting) is one of the most important components in many applications including multicriteria decision making, medicine, risk analysis, university ranking, submission decision in publication, information retrieval and filtering [1]. In these tasks, the attributes of the objects to be classified and the classes are ordered. An ordinal classifier f is expected to divide an unseen sample x into one of a set of ordered decision labels $D = \{\omega_1, \omega_2, \ldots, \omega_c\}$, according to the information provided with attributes $A = \{a_1, a_2, \ldots, a_m\}$, where an ordered relation exists between decision labels $\omega_1 < \omega_2 < \cdots < \omega_c$ like {High, Medium, Low} [2].

Compared with general classification problems, much less effort has been devoted to ordinal classification learning these years although this kind of tasks started to be discussed tens of years ago [3, 4]. In a number of literatures, the ordinal classification was transformed from a k-class ordinal problem to $k-1$ binary class problems [5, 6]. Then a learning algorithm for general classification tasks was employed on the derived data. In 2007, Cardoso and Costa [7] proposed a large-margin solution to ordinal classification, where the authors tempted to reduce k-class ordinal problem to k two-class problems and then large-margin classifiers were trained for these tasks. The same idea is also used in [8]. These techniques

*Corresponding author (email: huqinghua@hit.edu.cn) show to be effective in numerical experiments. Although they succeed in predicting new samples, they do not provide help in understanding the classification task as neural networks and SVMs are difficult to be understood for domain experts.

Decision tree induction is an efficient, effective and understandable technique for classification learning. Shannon's entropy plays a fundamental role in these algorithms including ID3 and C4.5, etc. These algorithms were generalized to address ordinal classification [9], however, it was proved that a training set in which all the examples are monotonic with respect to each other is not guaranteed to generate monotonic decision trees via information-theoretic top-down induction decision tree (TDIDT) algorithms that use entropy for attribute selection. The proof shows that Shannon entropy is inapplicable in the context of ordinal classification as it cannot measure the consistency of ordinal classification. In order to deal with this problem, a number of new measures were proposed. In [9], Ben-David introduced a non-monotonicity index defined as the ratio between the actual number of nonmonotonic branch pairs of a decision tree, and the maximum number of pairs that could have been non-monotonic with respect to each other in the same tree. In [10], an order-preserving tree-generation algorithm and an algorithm for repairing non-monotonic decision trees were provided for multi-attribute classification problems with k linearly ordered classes. In addition, some tree induction algorithms were also constructed to avoid violating the monotonicity of data [11, 12]. In 2008, Xia et al. [13] extended the Gini impurity used in CART to ordinal classification, and called it ranking impurity.

By replacing equivalence relations with dominance relations, Greco et al. [14] generalized the classical rough sets to dominance rough sets for analyzing multi-criterion decision making problems. Since then on, a collection of work has been reported to extend this model or use this model in various domains including fuzzy generalization [15], ordinal attribute reduction [16], university ranking [17] and assessment of bankruptcy risk [18]. In this model, the dependency function, which is the ratio of positive region over the universe, is considered as the measure of attribute quality. This ordinal dependency function really can capture and measure the consistency in ordinal classification tasks. However, like other dependency in rough set models this function is also sensitive to noisy information.

It is well known that the measure of mutual information derived from Shannon entropy outperforms the measures of Gini and dependency in decision tree construction [19–21]. In addition, mutual information also performs well in feature selection for evaluating quality of features [22], discretization for evaluating cutting sets [23], and registration of images [24], etc. This measure is stable and robust for sample perturbation and noise. It is desirable to develop a new information measure which holds the above advantages and is able to characterize the consistency of monotonicity in ordinal classification. In this work, we will introduce some new information measures for ordinal classification. These measures can be viewed as indexes of consistency between two rankings of a random variable with respect to different information; we call it ranking mutual information (RMI), while mutual information derived from Shannon entropy is a measure of classification consistency. The formulation of RMI is almost the same as Shannon's mutual information although they characterize different classes of relevance. Different from Spearman's rank correlation coefficient [25], RMI can be used to compute the relevance between two sets of variables, instead of two variables. Furthermore, we extend the RMI to the fuzzy context and show a measure of fuzzy ranking mutual information (FRMI). The proposed measures are used to evaluate the relevance between attributes and decision in ordinal classification.

The rest of the paper is organized as follows. Section 2 gives a review on Shannon's information measure in general classification. Section 3 introduces ranking entropy and ranking mutual information for ordinal classification. Section 4 extends ranking entropy and ranking mutual information in the fuzzy case. We compare the proposed measures with mutual information and dominance rough sets in section 5. Conclusions are presented in section 6.

2 Review on Shannon entropy in classification learning

Let U be the set of samples under consideration. $B \subseteq A$ is a subset of attributes and D is the decision. An equivalence relation R_B can be induced over U according to the values of samples on attributes B:

$R_B = \{(x_i, x_j)|\forall a \in B, a(x_i) = a(x_j)\}$, where $a(x)$ is the attribute value of sample x on a. Then a set of equivalence classes $\{X_1, X_2, \ldots, X_N\}$ are generated with the partition U/R_B, where the elements in X_i are indiscernible as their feature values are the same. Now we consider $X_1, X_2, \ldots, X_N$ are a set of random variables in U. The probability $p(X_i)$ of X_i is computed as $|X_i|/|U|$, then Shannon's entropy of the partition is defined as

$$H(B) = -\sum_{i=1}^{N} p(X_i)\log p(X_i), \tag{1}$$

where $|X_i|$ is the cardinality of set X_i.

Given another subset of attributes $C \subseteq A$, the partition induced by C is denoted by $\{Y_1, Y_2, \ldots, Y_M\}$, then the joint entropy of attributes B and C is computed as

$$H(B \cup C) = -\sum_{j=1}^{M}\sum_{i=1}^{N} p(X_i \cap Y_j)\log p(X_i \cap Y_j), \tag{2}$$

and the conditional entropy $H(B|C)$, reflecting the uncertainty of B if C is known, is defined as

$$H(B|C) = -\sum_{i=1}^{N}\sum_{j=1}^{M} p(X_i \cap Y_j)\log p(X_i|Y_j), \tag{3}$$

where $p(X_i|Y_j) = \frac{|X_i \cap Y_j|}{|Y_j|}$.

The mutual information MI of B and C is then defined as

$$MI(B, C) = -\sum_{i=1}^{N}\sum_{j}^{M} p(X_i \cap Y_j)\log \frac{p(X_i \cap Y_j)}{p(X_i)p(Y_i)}. \tag{4}$$

It is easy to get the following property that

$$MI(B, C) = MI(C, B) = H(B) - H(B|C) = H(C) - H(C|B).$$

The mutual information (MI) between features B and decision D: $MI(B, D) = -\sum_{i=1}^{N}\sum_{j}^{C} p(X_i \cap \omega_j) \log \frac{p(X_i \cap \omega_j)}{p(X_i)p(\omega_i)}$ characterizes the statistical relevance between B and D. This measure is widely discussed and used in feature selection and classification learning. In essence, MI reflects the degree of consistency of feature values and decision values of samples. If all the samples taken the same feature values are grouped into the same class, $MI(B, D)$ arrives at the maximal value $-\sum_{j}^{C} p(\omega_j)\log p(\omega_j)$ as for any X_i, we can find ω_j such that $X_i \subseteq \omega_j$.

In order to generalize Shannon's entropy in a natural way, we give a new formulation of the above measures. In the new formulation entropy, joint entropy, conditional entropy and mutual information are computed with the sum of uncertainty of single samples.

Let $[x_i]_B$ denote the equivalence class induced by sample x_i and attribute set B. The uncertainty of sample x_i is computed as

$$H_B(x_i) = -\log \frac{|[x_i]_B|}{|U|}, \tag{5}$$

and the average uncertainty of U, also called entropy of U with respect to B is computed as

$$H_B(U) = -\frac{1}{|U|}\sum_{i=1}^{n} \log \frac{|[x_i]_B|}{|U|}. \tag{6}$$

Now we give the definition of joint entropy:

$$H_{B \cup C}(U) = -\frac{1}{|U|}\sum_{i=1}^{n} \log \frac{|[x_i]_{B \cup C}|}{|U|} = -\frac{1}{|U|}\sum_{i=1}^{n} \log \frac{|[x_i]_B \cap [x_i]_C|}{|U|}, \tag{7}$$

conditional entropy

$$H_{B|C}(U) = H(B \cup C) - H(C) = -\frac{1}{|U|}\left(\sum_{i=1}^{n} \log \frac{|[x_i]_B \cap [x_i]_C|}{|U|} - \sum_{i=1}^{n} \log \frac{|[x_i]_C|}{|U|}\right)$$
$$= -\frac{1}{|U|}\sum_{i=1}^{n} \log \frac{|[x_i]_B \cap [x_i]_C|}{|[x_i]_C|}, \qquad (8)$$

and mutual information

$$H_{B,C}(U) = H(B) - H(B|C) = -\frac{1}{|U|}\sum_{i=1}^{n} \log \frac{|[x_i]_B|}{|U|} - \left(-\frac{1}{|U|}\sum_{i=1}^{n} \log \frac{|[x_i]_B \cap [x_i]_C|}{|[x_i]_C|}\right)$$
$$= -\frac{1}{|U|}\sum_{i=1}^{n}\left(\log \frac{|[x_i]_B|}{|U|} - \log \frac{|[x_i]_B \cap [x_i]_C|}{|[x_i]_C|}\right) = -\frac{1}{|U|}\sum_{i=1}^{n} \log \frac{|[x_i]_B| \times |[x_i]_C|}{|U||[x_i]_B \cap [x_i]_C|}. \qquad (9)$$

It is easy to show that

$$1) H(B) = H_B(U); \quad 2) H(B \cup C) = H_{B\cup C}(U); \quad 3) H(B|C) = H_{B|C}(U); \quad 4) H(B,C) = H_{B,C}(U).$$

We name formulae (5)–(9) as sample-wise information measures. With these formulae, it is easy to see the meaning of these measures. Given two features B and C, their mutual information reflects the overlap degree of samples with the same feature values. This degree should be kept in a general classification task, where the samples with the same feature values should be classified into the same decision class; otherwise, we think two decisions are not consistent. So the greater the overlap degree is, the more consistent the decision is.

3 Information measures for ordinal classification

The above analysis gives a new formulation of Shannon's information. These measures can be sample-wisely computed from the data. In addition, we also point out that MI in Shannon's entropy characterizes the consistency of classification, where the underlying assumption is the samples with the same feature values should be classified into the same decision; otherwise the classification is inconsistent. However, a different assumption of consistency is taken in ordinal classification; that is, the samples with the better feature values should not be grouped into a worse decision. This is, named as monotonicity constraints. Unfortunately MI in Shannon's theory cannot characterize this constraint. Now we introduce a new information measure for ordinal classification.

Let U be a set of samples described with a set of attributes A and a decision variable D. Given $\forall a \in A$ and $\forall x, y \in U$, we have $a(x) \leqslant a(y)$ or $a(x) \geqslant a(y)$. We say that x is better than y regarding $B \subseteq A$ if for $\forall a \in B$ we have $a(x) \geqslant a(y)$. We denote this relation by $x \geqslant_B y$. Similarly we can also denote $x \leqslant_B y$ if for $\forall a \in B$ we have $a(x) \leqslant a(y)$.

In addition, there is also a partially ordered structure between the decision labels $D = \{\omega_1, \omega_2, \ldots, \omega_c\}$ and we have $\omega_1 < \omega_2 < \cdots < \omega_c$.

Definition 1. A classification function f is said to be monotone with respect to B if

$$\text{for } x, y \in U : x \geqslant_B y \Rightarrow f(x) \geqslant f(y).$$

Definition 2. Given an ordinal classification sample set U described with a set of attributes A, for $\forall x \in U$, $B \subseteq A$, $a \in B$, we associate x with the following sets:

$$1)[x]_a^{\geqslant} = \{y \in U : y \geqslant_a x\}; \quad 2)[x]_B^{\geqslant} = \{y \in U : y \geqslant_B x\};$$
$$3)[x]_a^{\leqslant} = \{y \in U : y \leqslant_a x\}; \quad 4)[x]_B^{\leqslant} = \{y \in U : y \leqslant_B x\}.$$

It is easy to show $[x]_B^{\geqslant} = \bigcap\limits_{a\in B} [x]_a^{\geqslant}$, $[x]_B^{\leqslant} = \bigcap\limits_{a\in B} [x]_a^{\leqslant}$, $[x]_{B\cup C}^{\geqslant} = [x]_B^{\geqslant} \cap [x]_C^{\geqslant}$ and $[x]_{B\cup C}^{\leqslant} = [x]_B^{\leqslant} \cap [x]_C^{\leqslant}$.

Table 1 An ordinal classification task

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
a_1	1	1	3	2	2	3	3	4	5	5
a_2	1	2	2	3	3	3	4	4	4	5
D	1	1	1	2	2	2	2	3	3	3

Example 1. Given 10 manuscripts submitted to a journal, each is evaluated with two attributes: originality (a_1) and presentation (a_2), and a decision (D) is given to each manuscript. The samples are listed in Table 1.

Then we can compute that $[x_3]^{\geqslant}_{a_1} = \{x_3, x_6, x_7, x_8, x_9, x_{10}\}$; $[x_3]^{\leqslant}_{a_1} = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$; $[x_3]^{\geqslant}_{\{a_1,a_2\}} = \{x_3, x_6, x_7, x_8, x_9, x_{10}\}$; $[x_3]^{\leqslant}_{\{a_1,a_2\}} = \{x_1, x_2, x_3\}$.

Definition 3. Let U be a set of samples described with a set of attributes A, $B \subseteq A$. The upwards ranking entropy of the set U with respect to B is defined as

$$RH^{\geqslant}_{B}(U) = -\frac{1}{|U|}\sum_{i=1}^{n} \log \frac{|[x_i]^{\geqslant}_B|}{|U|}, \tag{10}$$

and downwards ranking entropy of the set U with respect to B is defined as

$$RH^{\leqslant}_{B}(U) = -\frac{1}{|U|}\sum_{i=1}^{n} \log \frac{|[x_i]^{\geqslant}_B|}{|U|}. \tag{11}$$

Example 2 (continue)**.** Given the information in Table 2, the ranking entropies of attributes a_1, a_2 and D are computed as follows:

$$\begin{aligned} RH^{\geqslant}_{\{a_1\}}(U) = & -\frac{1}{10}\sum_{i=1}^{10} \log \frac{|[x_i]^{\geqslant}_{\{a_1\}}|}{10} \\ = & -\frac{1}{10}\log\frac{10}{10} - \frac{1}{10}\log\frac{10}{10} - \frac{1}{10}\log\frac{6}{10} - \frac{1}{10}\log\frac{8}{10} - \frac{1}{10}\log\frac{8}{10} - \frac{1}{10}\log\frac{6}{10} \\ & - \frac{1}{10}\log\frac{6}{10} - \frac{1}{10}\log\frac{3}{10} - \frac{1}{10}\log\frac{2}{10} - \frac{1}{10}\log\frac{2}{10} \\ = & \ 0.9236. \end{aligned}$$

Analogically, $RH^{\geqslant}_{\{a_2\}}(U)$= 0.9135; $RH^{\geqslant}_{\{D\}}(U)$= 0.7269.

Property 1. We have $RH^{\geqslant}_{B}(U) \geqslant 0$ and $RH^{\leqslant}_{B}(U) \geqslant 0$ as $1 \geqslant \frac{|[x_i]^{\geqslant}_B|}{|U|} \geqslant 0$. $RH^{\leqslant}_{B}(U) = 0$ and $RH^{\geqslant}_{B}(U) = 0$ if and only if for $\forall x_i \in U$, $[x_i]^{\geqslant}_B = U$.

Proof. Straightforward.

Property 2. Let U be a set of samples described with a set of attributes A, $B \subseteq A$, $C \subseteq A$. If for $\forall x_i \in U$, $[x_i]^{\geqslant}_B \supseteq [x_i]^{\geqslant}_C$, $RH^{\geqslant}_{B}(U) \leqslant RH^{\geqslant}_{C}(U)$ or $[x_i]^{\leqslant}_B \supseteq [x_i]^{\leqslant}_C$, $RH^{\leqslant}_{B}(U) \leqslant RH^{\leqslant}_{C}(U)$.

Proof. Straightforward.

Corollary 1. Let U be a set of samples described with a set of attributes A, $B \subseteq C \subseteq A$. Then we have $RH^{\geqslant}_{B}(U) \leqslant RH^{\geqslant}_{C}(U)$ and $RH^{\leqslant}_{B}(U) \leqslant RH^{\leqslant}_{C}(U)$.

Proof. If $B \subseteq C$, we have $[x_i]^{\geqslant}_B \supseteq [x_i]^{\geqslant}_C$ and $[x_i]^{\leqslant}_B \supseteq [x_i]^{\leqslant}_C$. So $RH^{\geqslant}_{B}(U) \leqslant RH^{\geqslant}_{C}(U)$ and $RH^{\leqslant}_{B}(U) \leqslant RH^{\leqslant}_{C}(U)$.

Definition 4. Let U be a set of samples described with a set of attributes A, $B \subseteq A$, $C \subseteq A$. The upwards ranking joint entropy of the set U with respect to B and C is defined as

$$RH^{\geqslant}_{B\cup C}(U) = -\frac{1}{|U|}\sum_{i=1}^{n} \log \frac{|[x_i]^{\geqslant}_B \cap [x_i]^{\geqslant}_C|}{|U|}, \tag{12}$$

and downwards ranking joint entropy of the set U with respect to B and C is defined as

$$RH^{\leqslant}_{B\cup C}(U) = -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]^{\leqslant}_B \cap [x_i]^{\leqslant}_C|}{|U|}. \tag{13}$$

Corollary 2. Let U be a set of samples described with a set of attributes A, $B \subseteq A$, $C \subseteq A$. $RH^{\geqslant}_{B\cup C}(U) \geqslant RH^{\geqslant}_B(U)$; $RH^{\geqslant}_{B\cup C}(U) \geqslant RH^{\geqslant}_C(U)$; $RH^{\leqslant}_{B\cup C}(U) \geqslant RH^{\leqslant}_B(U)$; $RH^{\leqslant}_{B\cup C}(U) \geqslant RH^{\leqslant}_C(U)$.

Corollary 3. Let U be a set of samples described with a set of attributes A, $B \subseteq C \subseteq A$. Then we have $RH^{\geqslant}_{B\cup C}(U) = RH^{\geqslant}_C(U)$ and $RH^{\leqslant}_{B\cup C}(U) \leqslant RH^{\leqslant}_C(U)$.

Definition 5. Let U be a set of samples described with a set of attributes A, $B \subseteq A$, $C \subseteq A$. Known C, the upwards ranking conditional entropy of the set U with respect to B is defined as

$$RH^{\geqslant}_{B|C}(U) = -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]^{\geqslant}_B \cap [x_i]^{\geqslant}_C|}{|[x_i]^{\geqslant}_C|}, \tag{14}$$

and downwards ranking conditional entropy of the set U with respect to B is defined as

$$RH^{\leqslant}_{B|C}(U) = -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]^{\leqslant}_B \cap [x_i]^{\leqslant}_C|}{|[x_i]^{\geqslant}_C|}. \tag{15}$$

Property 3. Let U be a set of samples described with a set of attributes A, $B \subseteq A$ $C \subseteq A$. We have that $RH^{\leqslant}_{B|C}(U) = RH^{\leqslant}_{B\cup C}(U) - RH^{\leqslant}_C(U)$ and $RH^{\geqslant}_{B|C}(U) = RH^{\geqslant}_{B\cup C}(U) - RH^{\geqslant}_C(U)$.

Proof.

$$\begin{aligned}
RH^{\leqslant}_{B\cup C}(U) - RH^{\leqslant}_C(U) &= -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]^{\leqslant}_B \cap [x_i]^{\leqslant}_C|}{|U|} - \left(-\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]^{\leqslant}_C|}{|U|}\right)\\
&= -\frac{1}{|U|}\left(\sum_{i=1}^{n}\log\frac{|[x_i]^{\leqslant}_B \cap [x_i]^{\leqslant}_C|}{|U|} - \sum_{i=1}^{n}\log\frac{|[x_i]^{\leqslant}_C|}{|U|}\right)\\
&= -\frac{1}{|U|}\sum_{i=1}^{n}\left(\log\frac{|[x_i]^{\leqslant}_B \cap [x_i]^{\leqslant}_C|}{|U|} - \log\frac{|[x_i]^{\leqslant}_C|}{|U|}\right)\\
&= -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]^{\leqslant}_B \cap [x_i]^{\leqslant}_C|}{|[x_i]^{\leqslant}_C|}.
\end{aligned}$$

Similarly, we can also get that $RH^{\geqslant}_{B|C}(U) = RH^{\geqslant}_{B\cup C}(U) - RH^{\geqslant}_C(U)$.

Corollary 4. Let U be a set of samples described with a set of attributes A, $B \subseteq C \subseteq A$. We have that $RH^{\leqslant}_{B|C}(U) = 0$ and $RH^{\geqslant}_{B|C}(U) = 0$.

Proof. Assumed $B \subseteq C$, for $\forall x_i \in U$, $[x_i]^{\geqslant}_B \supseteq [x_i]^{\geqslant}_C$ and $[x_i]^{\leqslant}_B \supseteq [x_i]^{\leqslant}_C$. We have that $\frac{|[x_i]^{\leqslant}_B \cap [x_i]^{\leqslant}_C|}{|[x_i]^{\leqslant}_C|} = 1$.

$$RH^{\geqslant}_{B|C}(U) = -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]^{\geqslant}_B \cap [x_i]^{\geqslant}_C|}{|[x_i]^{\geqslant}_C|} = -\frac{1}{|U|}\sum_{i=1}^{n}\log 1 = 0.$$

Property 4. Let U be a set of samples described with a set of attributes A, $B \subseteq A$, $C \subseteq A$. We have that

1) $RH^{\leqslant}_{B\cup C}(U) \leqslant RH^{\leqslant}_B(U) + RH^{\leqslant}_C(U)$, $RH^{\geqslant}_{B\cup C}(U) \leqslant RH^{\geqslant}_B(U) + RH^{\geqslant}_C(U)$;

2) $RH^{\leqslant}_{B|C}(U) \leqslant RH^{\leqslant}_B(U), RH^{\leqslant}_{B|C}(U) \leqslant RH^{\leqslant}_C(U)$, $RH^{\geqslant}_{B|C}(U) \leqslant RH^{\geqslant}_B(U), RH^{\geqslant}_{B|C}(U) \leqslant RH^{\geqslant}_C(U)$.

Proof.

$$RH^{\leqslant}_{B\cup C}(U) - RH^{\leqslant}_B(U) - RH^{\leqslant}_C(U) = -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]^{\leqslant}_B \cap [x_i]^{\leqslant}_C|}{|U|}$$

$$
\begin{aligned}
&-\left(-\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]_B^{\leqslant}|}{|U|}-\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]_C^{\leqslant}|}{|U|}\right)\\
&=-\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]_B^{\leqslant}\cap[x_i]_C^{\leqslant}|}{|U|}+\left(\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]_B^{\leqslant}|}{|U|}\frac{|[x_i]_C^{\leqslant}|}{|U|}\right)\\
&=-\frac{1}{|U|}\sum_{i=1}^{n}\left(\log\frac{|[x_i]_B^{\leqslant}\cap[x_i]_C^{\leqslant}|}{|U|}-\log\frac{|[x_i]_B^{\leqslant}|}{|U|}\frac{|[x_i]_C^{\leqslant}|}{|U|}\right)\\
&=-\frac{1}{|U|}\sum_{i=1}^{n}\left(\log\frac{|[x_i]_B^{\leqslant}\cap[x_i]_C^{\leqslant}|\times|U|}{|[x_i]_B^{\leqslant}|\times|[x_i]_C^{\leqslant}|}\right).
\end{aligned}
$$

Definition 6. Let U be a set of samples described with a set of attributes A, $B\subseteq A$, $C\subseteq A$. The upwards ranking mutual information (URMI) of the set U regarding B and C is defined as

$$
RMI^{\geqslant}(B,C)=-\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]_B^{\geqslant}|\times|[x_i]_C^{\geqslant}|}{|U|\times|[x_i]_B^{\geqslant}\cap[x_i]_C^{\geqslant}|},\tag{16}
$$

and downwards ranking mutual information (DRMI) of the set U regarding B and C is defined as

$$
RMI^{\leqslant}(B,C)=-\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]_B^{\leqslant}|\times|[x_i]_C^{\leqslant}|}{|U|\times|[x_i]_B^{\leqslant}\cap[x_i]_C^{\leqslant}|}.\tag{17}
$$

In essence, ranking mutual information is the degree of monotonicity between features B and C. In ordinal classification, the monotonicity should be kept in classification learning. However, this structure of information is not considered in general classification tasks.

Ranking mutual information can be used to reflect the monotonicity relevance between features and decisions. So it is useful for ordinal feature selection and ordinal decision tree construction in ordinal classification, multicriteria decision making and ranking analysis.

Example 3 (continue). Given the information in Table 2, the ranking mutual information between a_1, a_2 and D are computed as follows:

$$
\begin{aligned}
RMI^{\geqslant}(\{a_1\},D)=&-\frac{1}{10}\sum_{i=1}^{10}\log\frac{|[x_i]_{\{a_1\}}^{\geqslant}|\times|[x_i]_D^{\geqslant}|}{10\times|[x_i]_{\{a_1\}}^{\geqslant}\cap[x_i]_D^{\geqslant}|}\\
=&-\frac{1}{10}\log\frac{10\times10}{10\times10}-\frac{1}{10}\log\frac{10\times10}{10\times10}-\frac{1}{10}\log\frac{6\times10}{10\times6}-\frac{1}{10}\log\frac{8\times7}{10\times7}-\frac{1}{10}\log\frac{8\times7}{10\times7}\\
&-\frac{1}{10}\log\frac{6\times7}{10\times5}-\frac{1}{10}\log\frac{6\times7}{10\times5}-\frac{1}{10}\log\frac{3\times3}{10\times3}-\frac{1}{10}\log\frac{2\times3}{10\times2}-\frac{1}{10}\log\frac{2\times3}{10\times2}\\
=&\,0.6358.
\end{aligned}
$$

Analogically, $RMI^{\geqslant}(\{a_2\},D)$= 0.6439.

Property 5. Let U be a set of samples described with a set of attributes A, $B\subseteq A$, $C\subseteq A$. We have that

1) $RMI^{\leqslant}(B,C)=RH_B^{\leqslant}(U)-RH_{B|C}^{\leqslant}(U)=RH_C^{\leqslant}(U)-RH_{C|B}^{\leqslant}(U)$,

2) $RMI^{\geqslant}(B,C)=RH_B^{\geqslant}(U)-RH_{B|C}^{\geqslant}(U)=RH_C^{\geqslant}(U)-RH_{C|B}^{\geqslant}(U)$.

Proof.

$$
\begin{aligned}
RH_B^{\leqslant}(U)-RH_{B|C}^{\leqslant}(U)=&-\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]_B^{\geqslant}|}{|U|}-\left(-\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]_B^{\geqslant}\cap[x_i]_C^{\geqslant}|}{|[x_i]_C^{\geqslant}|}\right)\\
=&-\frac{1}{|U|}\left(\sum_{i=1}^{n}\log\frac{|[x_i]_B^{\geqslant}|}{|U|}-\sum_{i=1}^{n}\log\frac{|[x_i]_B^{\geqslant}\cap[x_i]_C^{\geqslant}|}{|[x_i]_C^{\geqslant}|}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{1}{|U|}\sum_{i=1}^{n}\left(\log\frac{|[x_i]_B^{\geqslant}|}{|U|} - \log\frac{|[x_i]_B^{\geqslant}\cap[x_i]_C^{\geqslant}|}{|[x_i]_C^{\geqslant}|}\right)\\
&= -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]_B^{\geqslant}|\times|[x_i]_C^{\geqslant}|}{|U|\times|[x_i]_B^{\geqslant}\cap[x_i]_C^{\geqslant}|} = RMI^{\leqslant}(B,C).
\end{aligned}
$$

In the same way, we can also derive that $RMI^{\leqslant}(B,C) = RH_C^{\leqslant}(U) - RH_{C|B}^{\leqslant}(U)$, $RMI^{\geqslant}(B,C) = RH_B^{\geqslant}(U) - RH_{B|C}^{\geqslant}(U) = RH_C^{\geqslant}(U) - RH_{C|B}^{\geqslant}(U)$.

Property 6. Let U be a set of samples described with a set of attributes A and $B \subseteq C \subseteq A$. We have that

1) $RMI^{\geqslant}(B,C) = RH_B^{\geqslant}(U)$;
2) $RMI^{\leqslant}(B,C) = RH_B^{\leqslant}(U)$.

Proof. $RMI^{\geqslant}(B,C) = -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]_B^{\geqslant}|\times|[x_i]_C^{\geqslant}|}{|U|\times|[x_i]_B^{\geqslant}\cap[x_i]_C^{\geqslant}|}$. If $B \subseteq C$, we have $[x_i]_B^{\geqslant} \supseteq [x_i]_C^{\geqslant}$. So $[x_i]_B^{\geqslant} \cap [x_i]_C^{\geqslant} = [x_i]_C^{\geqslant}$.

In this case $RMI^{\geqslant}(B,C) = -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]_B^{\geqslant}|}{|U|} = RH_B^{\geqslant}(U)$. Analogously, we can also derive that $RMI^{\leqslant}(B,C) = RH_B^{\leqslant}(U)$.

Corollary 5. Let U be a set of samples described with a set of attributes A, D is the decision attribute and $B \subseteq A$. If for $\forall x_i \in U$ we have $[x_i]_B^{\geqslant} \subseteq [x_i]_D^{\geqslant}$, then we say that the decision is upwards consistent and $RMI^{\geqslant}(B,D) = RH_D^{\geqslant}(U)$. If for $\forall x_i \in U$ we have that $[x_i]_B^{\leqslant} \subseteq [x_i]_D^{\leqslant}$, then we say the decision is downwards consistent and $RMI^{\leqslant}(B,C) = RH_D^{\leqslant}(U)$.

Proof. Straightforward.

Given the information in Table 2, we get that $[x_8]_{a_2}^{\geqslant} \not\subseteq [x_8]_D^{\geqslant}$, $[x_9]_{a_2}^{\geqslant} \not\subseteq [x_9]_D^{\geqslant}$, so the decision is upwards inconsistent. In this case $RMI^{\geqslant}(\{a_2\},D) < RH_D^{\geqslant}(U)$.

4 Information measures in fuzzy ordinal classification

Let U be a set of samples described with numerical or fuzzy features and $x, y \in U$. Sometimes one wants to know not only whether $a(x) \leqslant a(y)$ or $a(x) \geqslant a(y)$, but also how much x is greater than or less than y. Fuzzy ranking can be introduced to represent this kind of structures. The above measures are not applicable in this case. Now we extend these measures into the fuzzy case.

Example 4. Assume the 10 manuscripts in Table 1 are evaluated with two attributes: originality (a_1) and presentation (a_2), and a decision (D) is given to each manuscript, as shown in Table 2.

According to the information in Table 2, we can also compute the ranking of these samples with respect to features a_1 and a_2. For example, $[x_3]_{a_1}^{\geqslant} = \{x_3, x_7, x_8, x_9, x_{10}\}$ and $[x_3]_{a_1}^{\geqslant} = \{x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$. In this case, we use a crisp membership to describe the subset of samples which are not less than x_3 in terms of a_1 and a_2, respectively. The membership function is shown as Figure 1(a).

Figure 1(a) describes a crisp set of objects greater than 0.5, while Figure 1(b) gives a fuzzy set greater than 0.5. Now we introduce a function to compute fuzzy ordinal set:

$$
f(x) = \frac{1}{1 + \mathrm{e}^{-k[a(x)-a(y)]}},
$$

where k is a parameter to regulate the preference degree by users.

Table 2 An ordinal classification task

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
a_1	0.28	0.25	0.60	0.48	0.42	0.55	0.78	0.75	0.83	0.85
a_2	0.28	0.31	0.42	0.47	0.51	0.58	0.71	0.78	0.80	0.91
D	1	1	1	2	2	2	2	3	3	3

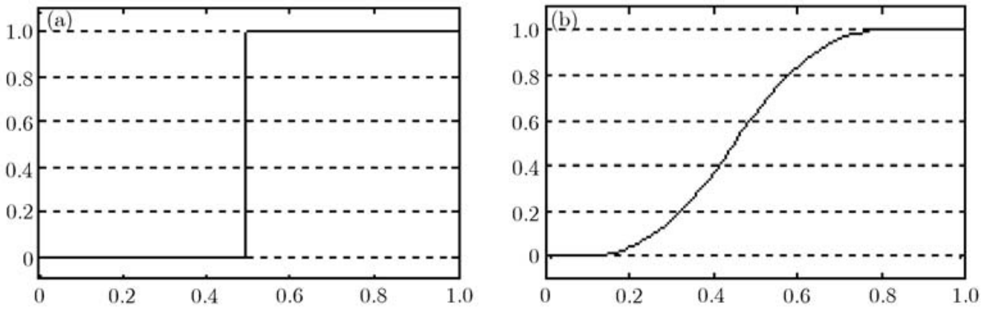

Figure 1 Membership functions of ordinal sets. (a) Crisp ordinal set; (b) fuzzy ordinal set.

Certainly, there are a number of functions to be used in computing fuzzy ranking. Now the fuzzy set greater than or less than x_i with respect to feature a can be denoted by

$$\widetilde{[x_i]_a^{\geqslant}} = \frac{r_{1i}}{x_1} + \frac{r_{2i}}{x_2} + \cdots + \frac{r_{ji}}{x_j} + \cdots + \frac{r_{ni}}{x_n} \quad \text{or} \quad \widetilde{[x_i]_a^{\leqslant}} = \frac{s_{1i}}{x_1} + \frac{s_{2i}}{x_2} + \cdots + \frac{s_{ji}}{x_j} + \cdots + \frac{s_{ni}}{x_n},$$

where $r_{ji} = \frac{1}{1+\mathrm{e}^{-k[a(x_j)-a(x_i)]}}$, $s_{ji} = \frac{1}{1+\mathrm{e}^{-k[a(x_i)-a(x_j)]}}$ and $k > 0$.

Let U be a set of samples described with a set of attributes A, $a \in A$, $b \in A$. Assumed that $\widetilde{[x_i]_a^{\geqslant}} = \frac{r_{1i}}{x_1} + \frac{r_{2i}}{x_2} + \cdots + \frac{r_{ji}}{x_j} + \cdots + \frac{r_{ni}}{x_n}$, $\widetilde{[x_i]_b^{\geqslant}} = \frac{s_{1i}}{x_1} + \frac{s_{2i}}{x_2} + \cdots + \frac{s_{ji}}{x_j} + \cdots + \frac{s_{ni}}{x_n}$, we define that

$$\widetilde{[x_i]^{\geqslant}_{\{a\}\cup\{b\}}} = \frac{\min(r_{1i}, s_{1i})}{x_1} + \frac{\min(r_{2i}, s_{2i})}{x_2} + \cdots + \frac{\min(r_{ji}, s_{ji})}{x_j} + \cdots + \frac{\min(r_{ni}, s_{ni})}{x_n}.$$

Similarly, if we know $\widetilde{[x_i]_a^{\leqslant}}$ and $\widetilde{[x_i]_b^{\leqslant}}$, we can compute $\widetilde{[x_i]^{\leqslant}_{\{a\}\cup\{b\}}}$. In addition, if $B = \{a_1, a_2, \ldots, a_k\}$, we get that $\widetilde{[x_i]_B^{\geqslant}} = \widetilde{[x_i]^{\geqslant}_{\{a_1\}\cup\{a_2\}\cup\cdots\cup\{a_k\}}}$.

Simulating the definition in [26, 27], we give fuzzy ranking entropy and fuzzy ranking mutual information as follows.

Definition 7. Let U be a set of samples described with a set of attributes A, $B \subseteq A$. For $\forall x_i \in U$, $\widetilde{[x_i]_B^{\geqslant}} = \frac{r_{1i}}{x_1} + \frac{r_{2i}}{x_2} + \cdots + \frac{r_{ji}}{x_j} + \cdots + \frac{r_{ni}}{x_n}$ and $\widetilde{[x_i]_B^{\leqslant}} = \frac{s_{1i}}{x_1} + \frac{s_{2i}}{x_2} + \cdots + \frac{s_{ji}}{x_j} + \cdots + \frac{s_{ni}}{x_n}$. Then the upwards fuzzy ranking entropy of the set U with respect to B is defined as

$$FRH_B^{\geqslant}(U) = -\frac{1}{|U|}\sum_{i=1}^{n} \log \frac{|\widetilde{[x_i]_B^{\geqslant}}|}{|U|}, \tag{18}$$

and downwards fuzzy ranking entropy of the set U with respect to B is defined as

$$FRH_B^{\leqslant}(U) = -\frac{1}{|U|}\sum_{i=1}^{n} \log \frac{|\widetilde{[x_i]_B^{\leqslant}}|}{|U|}, \tag{19}$$

where $|\widetilde{[x_i]_B^{\geqslant}}| = \sum_{j=1}^{n} r_{jn}$ and $|\widetilde{[x_i]_B^{\leqslant}}| = \sum_{j=1}^{n} s_{jn}$ are the fuzzy cardinality of these fuzzy sets.

Definition 8. Let U be a set of samples described with a set of attributes A, $B \subseteq A$, $C \subseteq A$. Then the upwards fuzzy ranking joint entropy of the set U with respect to B and C is defined as

$$FRH_{B\cup C}^{\geqslant}(U) = -\frac{1}{|U|}\sum_{i=1}^{n} \log \frac{|\widetilde{[x_i]_B^{\geqslant}} \cap \widetilde{[x_i]_C^{\geqslant}}|}{|U|} = -\frac{1}{|U|}\sum_{i=1}^{n} \log \frac{|\widetilde{[x_i]_{B\cup C}^{\geqslant}}|}{|U|}, \tag{20}$$

and downwards fuzzy ranking joint entropy of the set U with respect to B and C is defined as

$$FRH_{B\cup C}^{\leqslant}(U) = -\frac{1}{|U|}\sum_{i=1}^{n} \log \frac{|\widetilde{[x_i]_B^{\leqslant}} \cap \widetilde{[x_i]_C^{\leqslant}}|}{|U|} = -\frac{1}{|U|}\sum_{i=1}^{n} \log \frac{|\widetilde{[x_i]_{B\cup C}^{\leqslant}}|}{|U|}. \tag{21}$$

Definition 9. Let U be a set of samples described with a set of attributes A, $B \subseteq A$, $C \subseteq A$. Then the upwards fuzzy ranking conditional entropy of the set U with respect to B in the case that C is known is defined as

$$FRH^{\geqslant}_{B\cup C}(U) = -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|\widetilde{[x_i]^{\geqslant}_B} \cap \widetilde{[x_i]^{\geqslant}_C}|}{|\widetilde{[x_i]^{\geqslant}_C}|} = -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|\widetilde{[x_i]^{\geqslant}_{B\cup C}}|}{|\widetilde{[x_i]^{\geqslant}_C}|}, \tag{22}$$

and downwards fuzzy ranking conditional entropy of the set U with respect to B in the case that C is known is defined as

$$FRH^{\leqslant}_{B\cup C}(U) = -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|\widetilde{[x_i]^{\leqslant}_B} \cap \widetilde{[x_i]^{\leqslant}_C}|}{|\widetilde{[x_i]^{\leqslant}_C}|} = -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|\widetilde{[x_i]^{\leqslant}_{B\cup C}}|}{|\widetilde{[x_i]^{\leqslant}_C}|}. \tag{23}$$

Definition 10. Let U be a set of samples described with a set of attributes A, $B \subseteq A$, $C \subseteq A$. Then the upwards fuzzy ranking mutual information of the set U with respect to B and C is defined as

$$FRMI^{>}(B,C) = -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|\widetilde{[x_i]^{\geqslant}_B}| \times |\widetilde{[x_i]^{\geqslant}_C}|}{|U| \times |\widetilde{[x_i]^{\geqslant}_B} \cap \widetilde{[x_i]^{\geqslant}_C}|}, \tag{24}$$

and the downwards fuzzy ranking mutual information of the set U with respect to B and C is defined as

$$FRMI^{<}(B,C) = -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|\widetilde{[x_i]^{\leqslant}_B}| \times |\widetilde{[x_i]^{\leqslant}_C}|}{|U| \times |\widetilde{[x_i]^{\leqslant}_B} \cap \widetilde{[x_i]^{\leqslant}_C}|}. \tag{25}$$

The properties of fuzzy ranking entropy, joint entropy, conditional entropy and mutual information are the same as the crisp one. We do not discuss them here.

5 Comparison of ranking mutual information with related measures

There are a number of techniques to compute relevance between variables. In this section, we will compare the ranking mutual information with two important methods: mutual information and dominance rough sets.

5.1 Comparison of ranking mutual information with mutual information

Mutual information is widely used in measuring relevance between random variables in classification learning [28, 29]. The assumption of consistency in classification is different from that in ranking. One expects the samples with the identical feature values should be grouped into a decision. Mutual information reflects the degree that how many samples with the same feature values are divided into the same decisions. Therefore, mutual information does not care the order of feature values, but their equivalence.

Assume we get a set of samples, as shown in Table 3. There are 12 samples characterized with one attribute and one decision variable.

As to mutual information, the rank is not cared. In this case, we can see that the decision is consistent because all samples with the same feature values are grouped into the same decision classes. We can get four consistent decision rules: (1) if a=0, then D=4; (2) if a=1 or 2 or 6, then D=4; (3) if a=3, then D=2; (4) if a=4 or 5, then D=3.

We compute the mutual information between a and D.

$$\begin{aligned}
H_{\{a\},D}(U) = & -\frac{1}{|U|}\sum_{i=0}^{11}\log\frac{|[x_i]_a| \times |[x_i]_D|}{|U||[x_i]_a \cap [x_i]_D|} = -\frac{1}{12}\sum_{i=0}^{11}\log\frac{|[x_i]_a| \times |[x_i]_D|}{12 \times |[x_i]_a \cap [x_i]_D|} \\
= & -\frac{1}{12}\log\frac{|\{x_0\}| \times |\{x_0\}|}{12 \times |\{x_0\}|} - \frac{1}{12}\log\frac{|\{x_1\}| \times |\{x_1, x_2, x_3, x_{11}\}|}{12 \times |\{x_1\}|} - \cdots \\
& -\frac{1}{12}\log\frac{|\{x_{11}\}| \times |\{x_1, x_2, x_3, x_{11}\}|}{12 \times |\{x_{11}\}|} = 1.8554
\end{aligned}$$

Table 3 An ordinal classification task

	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
a	0	1	2	2	3	3	3	4	4	4	5	6
D	4	1	1	1	2	2	2	3	3	3	3	1

$$\begin{aligned} H(D) = & -(1/12)\times\log(1/12)-(4/12)\times\log(4/12)-(3/12)\times\log(3/12) \\ & -(4/12)\times\log(4/12)=1.8554. \end{aligned}$$

$H_{\{a\},D}(U)=H(D)$ shows the information provided by attribute a equals to the uncertainty quantity of decision D. These samples can consistently be discriminated according to the information provided by attribute a.

If we consider the order of these feature values. The conclusion is completely different. Object x_0 with the worst feature value 0 gets the best decision 4, whereas object x_{11} with the best feature value 6 gets the worst decision 1. Obviously, the ranking is not consistent. Mutual information cannot characterize this kind of inconsistency.

We also compute the ranking mutual information between a and D as follows:

$$\begin{aligned} RMI^{\geqslant}(\{a\},D) = & -\frac{1}{12}\sum_{i=0}^{11}\log\frac{|[x_i]_{\{a\}}^{\geqslant}|\times|[x_i]_D^{\geqslant}|}{|U|\times|[x_i]_{\{a\}}^{\geqslant}\cap[x_i]_D^{\geqslant}|}=-\frac{1}{12}\log\frac{12\times 1}{12\times 1}-\frac{1}{12}\log\frac{11\times 12}{12\times 11} \\ & -\frac{2}{12}\log\frac{10\times 12}{12\times 10}-\frac{3}{12}\log\frac{8\times 8}{12\times 7}-\frac{3}{12}\log\frac{5\times 5}{12\times 4} \\ & -\frac{1}{12}\log\frac{2\times 2}{12\times 1}-\frac{1}{12}\log\frac{1\times 12}{12\times 1}=0.4654, \end{aligned}$$

$$RH_D^{\geqslant}(U)=-\frac{1}{12}\sum_{i=1}^{n}\log\frac{|[x_i]_B^{\geqslant}|}{12}=-\frac{1}{12}\log\frac{1}{12}-\frac{4}{12}\log\frac{12}{12}-\frac{3}{12}\log\frac{9}{12}-\frac{4}{12}\log\frac{5}{12}=0.8235$$

The ranking mutual information between a and D is less than the uncertainty quantity of decision D. The ranking is not consistent. Correspondingly we cannot get monotonic decision rules in this case.

5.2 Comparison of ranking mutual information with dominance rough sets

Dominance rough sets, extended from Pawlak's rough set [30], have the ability to reflect the monotonicity in samples by using dominance relation in approximation [14, 31].

Let U be a set of samples described with attributes A and a decision D. The decision labels $D=\{\omega_1,\omega_2,\ldots,\omega_c\}$ and we have $\omega_1<\omega_2<\cdots<\omega_c$. We denote the subsets of samples which get no worse or no better decision than ω_i by $Cl_i^{\geqslant}=\cup_{s\geqslant i}\omega_i$ and $Cl_i^{\leqslant}=\cup_{s\leqslant i}\omega_i$, respectively. Then we use sample subsets $[x]_B^{\geqslant}$ or $[x]_B^{\leqslant}$ to approximate the $Cl_i^{\geqslant}$ and $Cl_i^{\leqslant}$. Given attributes B, the lower and upper approximations of $Cl_i^{\geqslant}$ and $Cl_i^{\leqslant}$ are defined as

$$\begin{aligned} &\underline{B}(Cl_i^{\geqslant})=\{x\in U|[x]_B^{\geqslant}\subseteq Cl_i^{\geqslant}\}, \quad \overline{B}(Cl_i^{\geqslant})=\{x\in U|[x]_B^{\leqslant}\cap Cl_i^{\geqslant}\neq\emptyset\}, \\ &\underline{B}(Cl_i^{\leqslant})=\{x\in U|[x]_B^{\leqslant}\subseteq Cl_i^{\leqslant}\}, \quad \overline{B}(Cl_i^{\leqslant})=\{x\in U|[x]_B^{\geqslant}\cap Cl_i^{\leqslant}\neq\emptyset\}. \end{aligned}$$

Corresponding, the approximation boundary region is computed with

$$BN_B(Cl_i^{\geqslant})=\overline{B}(Cl_i^{\geqslant})-\underline{B}(Cl_i^{\geqslant}), \quad BN_B(Cl_i^{\leqslant})=\overline{B}(Cl_i^{\leqslant})-\underline{B}(Cl_i^{\leqslant}).$$

It is easy to get that $BN_B(Cl_i^{\geqslant})=BN_B(Cl_{i-1}^{\leqslant})$ [31]. As we know, the decisions of boundary samples are not consistent as they violate the principle of monotonicity. The approximation quality, also called dependency of D on B is defined as

$$\gamma_B(D)=\frac{|U-\overset{c}{\underset{i=1}{\cup}}BN_B(Cl_i^{\geqslant})|}{|U|}=\frac{|U-\overset{c}{\underset{i=1}{\cup}}BN_B(Cl_i^{\leqslant})|}{|U|}.$$

Dependency $\gamma_B(D)$ is the ratio of consistent samples over the whole samples. Therefore, dependency can be considered as a measure of ranking relevance between B and D. What is the difference between ranking mutual information and dependency in dominance rough sets?

Observing $RMI^{\geqslant}(B,D) = -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]_B^{\geqslant}|\times|[x_i]_D^{\geqslant}|}{|U|\times|[x_i]_B^{\geqslant}\cap[x_i]_D^{\geqslant}|}$ and $\gamma_B(D) = \frac{|U-\cup_{i=1}^{c}BN_B(Cl_i^{\geqslant})|}{|U|}$, we can see that if $x \in U$ is consistent, namely $[x]_B^{\geqslant} \subseteq [x_i]_D^{\geqslant}$, sample x contributes $1/|U|$ to dependency and contributes $-\frac{1}{|U|}\log\frac{|[x_i]_D^{\geqslant}|}{|U|}$ to ranking mutual information. So $\gamma_B(D) = |U|/|U| = 1$ and $RMI(B,D) = -\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|[x_i]_D^{\geqslant}|}{|U|} = RH^{\geqslant}(D)$ if all samples are consistent. However, if sample x is not consistent with respect to attribute B. In this case, x belongs to the boundary region of decision. Sample x contributes nothing to dependency and contributes $-\frac{1}{|U|}\log\frac{|[x_i]_B^{\geqslant}|\times|[x_i]_D^{\geqslant}|}{|U|\times|[x_i]_B^{\geqslant}\cap[x_i]_D^{\geqslant}|}$ to ranking mutual information. As $\frac{|[x_i]_B^{\geqslant}|\times|[x_i]_D^{\geqslant}|}{|U|\times|[x_i]_B^{\geqslant}\cap[x_i]_D^{\geqslant}|} \geqslant \frac{|[x_i]_D|}{|U|}$, so this contribution is no more than that if x is consistent. We can also get the same result for $RMI^{\leqslant}(B,D)$.

The analysis shows that inconsistent samples may have contribution to ranking mutual information, while have no contribution to dependency. Due to this property, ranking mutual information and fuzzy ranking mutual information are more robust to noisy samples than dependency.

Let us look the samples in Table 3. If we do not consider x_0 and x_{11}, it is easy to get that $\gamma_{\{a\}}(D) = 1$ and

$$\begin{aligned} RMI^{\geqslant}(\{a\},D) = & -\frac{1}{10}\log\frac{10\times 10}{10\times 10} - \frac{2}{10}\log\frac{9\times 10}{10\times 9} - \frac{3}{10}\log\frac{7\times 7}{10\times 7} - \frac{3}{10}\log\frac{4\times 4}{10\times 4} - \frac{1}{10}\log\frac{1\times 4}{10\times 1} \\ & = 0.6831. \end{aligned}$$

However, if there are two noisy samples x_0 and x_{11} in the set of samples. Then there are just three consistent samples x_1, x_2, x_3. The dependency is $\gamma_{\{a\}}(D) = 3/12 = 0.25$. Moreover, if the decision of x_{11} is 0, then $\gamma_{\{a\}}(D) = 0$. Therefore, dependency is very sensitive to the noisy samples. Now we compute ranking mutual information in presence of noisy samples x_0 and x_{11}, and the decision of x_{11} is 0. In this case, $RMI_1^{\geqslant}(\{a\},D) = 0.9397$. Assume that we know the decisions of x_0 and x_{11} are mislabeled. Their real decisions are 0 and 4, respectively. Then we get that $RMI_2^{\geqslant}(\{a\},D) = 0.9544$. We can see that the difference between $RMI_1^{\geqslant}$ and $RMI_2^{\geqslant}$ is very little. This discussion shows that ranking mutual information is more robust than dependency if some noisy samples exist.

6 Conclusions

Ordinal classification is a class of important learning tasks in decision analysis. Compared with general classification, little effort has been devoted to construct learning algorithms for this kind of tasks. Shannon information entropy and the derived measure of mutual information play a fundamental role in a set of learning algorithms. However, these measures are not applicable to ordinal classification.

In this work, we reform Shanon's entropy and mutual information with a sample-wise formulation. Then we extend these measures into the context of ordinal classification and fuzzy ordinal classification. We discuss the properties of these measures and show that the proposed ranking mutual information and fuzzy ranking mutual information are the indexes of consistency of monotonicity in ordinal classification. Thus, these measures can be used to evaluate features and select features in the case of ordinal classification.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 60703013, 10978011), the Key Program of the National Natural Science Foundation of China (Grant No. 60932008), the National Science Fund for Distinguished Young Scholars (Grant No. 50925625), and China Postdoctoral Science Foundation.

References

1 Kamishima T, Akaho S. Dimension reduction for supervised ordering. In: Proceedings of the Sixth International Conference on Data Mining (ICDM'06). Hong Kong, China, 2006. 18–22

2 Lee J W T, Yeung D S, Wang X. Monotonic decision tree for ordinal classification. IEEE Int Conf Syst Man Cybern, 2003, 3: 2623–2628

3 Ben-David A, Sterling L, Pao Y H. Learning and classification of monotonic ordinal concepts. Comput Intell, 1989, 5: 45–49

4 Ben-David A. Automatic generation of symbolic multiattribute ordinal knowledge-based DSSs: Methodology and applications. Decis Sci, 1992, 23: 1357–1372

5 Frank E, Hall M. A simple approach to ordinal classification. In: De Raedt L, Flach P, eds. ECML 2001, LNAI 2167. Berlin: Springer-Verlag, 2001. 145–156

6 Costa J P, Cardoso J S. Classification of ordinal data using neural networks. In: Gama J, Camacho R, Brazdil P, et al. eds. ECML 2005, LNAI 3720. Berlin: Springer-Verlag, 2005. 690–697

7 Cardoso J S, Costa J F P. Learning to classify ordinal data: the data replication method. J Mach Learn Res, 2007, 8: 1393–1429

8 Costa J P, Alonso H, Cardoso J S. The unimodal model for the classification of ordinal data. Neur Netw, 2008, 21: 78–91

9 Ben-David A. Monotonicity maintenance in information-theoretic machine learning algorithms. Mach Learn, 1995, 19: 29–43

10 Potharst R, Bioch J C. Decision trees for ordinal classification. Intell Data Anal, 2000, 4: 97–111

11 Cao-Van K, Baets B D. Growing decision trees in an ordinal setting. Int J Intell Syst, 2003, 18: 733–750

12 Potharst R, Feelders A J. Classification trees for problems with monotonicity constraints. ACM SIGKDD Explor Newslett, 2002, 4: 1–10

13 Xia F, Zhang W S, Li F X, et al. Ranking with decision tree. Know Inf Syst, 2008, 17: 381–395

14 Greco S, Matarazzo B, Slowinski R. Rough approximation of a preference relation by dominance relations. ICS Research Report 16/96. Europ J Operat Res, 1999, 117: 63–83

15 Hu Q, Yu D, Guo M Z. Fuzzy preference based rough sets. Inf Sci, 2010, 180: 2003–2022

16 Lee J W T, Yeung D S, Tsang E C C. Rough sets and ordinal reducts. Soft Comput, 2006, 10: 27–33

17 Sai Y, Yao Y Y, Zhong N. Data analysis and mining in ordered information tables. In: Proceedings of the IEEE International Conference on Data Mining, IEEE Computer Society, 2001. 497–504

18 Greco S, Matarazzo B, Slowinski R. Rough sets methodology for sorting problems in presence of multiple attributes and criteria. Europ J Operat Res, 2002, 138: 247–259

19 Liang J Y, Qian Y H. Information granules and entropy theory in information systems. Sci China Ser F-Inf Sci, 2008, 51: 1427–1444

20 Hu D, Li H X, Yu X C. The information content of rules and rule sets and its application. Sci China Ser F-Inf Sci, 2008, 51: 1958–1979

21 Mingers J. An empirical comparison of selection measures for decision-tree induction. Mach Learn, 1989, 3: 319–342

22 Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Patt Anal Mach Intell, 2005, 27: 1226–1238

23 Fayyad U M, Irani K B. On the handling of continuous-valued attributes in decision tree generation. Mach Learn, 1992, 8: 87–102

24 Viola P, Wells W M. III. Alignment by maximization of mutual information. Int J Comput Vision, 1997, 24: 137–154

25 Spearman C. "Footrule" for measuring correlation. British J Psych, 1906, 2: 89–108

26 Hu Q H, Yu D R, Xie Z X, et al. Fuzzy probabilistic approximation spaces and their information measures. IEEE Trans Fuzzy Syst, 2006, 14 : 191–201

27 Yu D R, Hu Q H, Wu C. Uncertainty measures for fuzzy relations and their applications. Appl Soft Comput, 2007, 7: 1135–1143

28 Quinlan J R. Induction of decision trees. Mach Learn 1986, 1: 81–106

29 Quinlan J R. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993

30 Pawlak Z. Rough Sets, Theoretical Aspects of Reasoning About Data. Dordrecht: Kluwer Academic Publishers, 1991

31 Greco S, Matarazzo B, Slowinski R. Rough approximation by dominance relations. Int J Intell Syst, 2002, 17: 153–171