Entropy **2013**, *15*, 2288-2302; doi:10.3390/e15062288

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Multi-Granulation Entropy and Its Applications

Kai Zeng *, Kun She and Xinzheng Niu

School of Computer Science & Engineering, University of Electronic Science and Technology of China, Sichuan, Chengdu 611731, China; E-Mails: kunshe@126.com (K.S.); xinzhengniu@hotmail.com (X.Z.N.)

* Author to whom correspondence should be addressed; E-Mail: zengkailink@sina.com.

Received: 2 April 2013; in revised form: 22 May 2013 / Accepted: 30 May 2013 /
Published: 6 June 2013

Abstract: In the view of granular computing, some general uncertainty measures are proposed through single-granulation by generalizing Shannon's entropy. However, in the practical environment we need to describe concurrently a target concept through multiple binary relations. In this paper, we extend the classical information entropy model to a multi-granulation entropy model (MGE) by using a series of general binary relations. Two types of MGE are discussed. Moreover, a number of theorems are obtained. It can be concluded that the single-granulation entropy is the special instance of MGE. We employ the proposed model to evaluate the significance of the attributes for classification. A forward greedy search algorithm for feature selection is constructed. The experimental results show that the proposed method presents an effective solution for feature analysis.

Keywords: multi-granulation; entropy; feature selection

1. Introduction

Uncertainty analysis represents one of the most significant challenging tasks in intelligent computation. Since Shannon introduced the information entropy to measure the uncertainty of the system, a series of measures were proposed for machine learning, data mining and pattern recognition, *etc.* [1–3].

In the field of granular computing, Yu *et al.* introduced the fuzzy entropy for attribute reduction [4]. Hu *et al.* presented kernel entropy by extended Yu's work [5]. In [6], the authors defined neighborhood entropy by using a neighborhood relation. In the view of granular computing, there are two modules in the entropy methodology mentioned above: (1) granulation of data (samples) into a set of information granules according to the relation of objects; (2) calculating the sum of the uncertainty quantity of all the information granules. We will give an example to illustrate this two-step process in detail in Section 2. It shows that granulation plays a key role in these entropy models. However, the classical information entropy theory utilizes solely the granularity structure of the given data, which is expressed by one suitable binary relation. The neighborhood entropy is only based on the neighborhood granulation; the fuzzy entropy on the fuzzy granulation; and the kernel entropy on the kernel granulation. In [7], Qian *at el.* proposed that there is a contradiction between two different binary relations in some data analysis issues. In other words, the decision or the view of each of decision makers may be independent for the same object in the process of some decision making. Accordingly, Qian *et al.* proposed multi-granulation rough set (MGRS) according to a user's different requirements or targets of problem solving. Since then, many researchers have extended the classical MGRS by using various generalized binary relations. Lin *et al.* [8] proposed a covering-based pessimistic multi-granulation rough set. Xu *et al.* [9] proposed another generalized version, called variable precision multi-granulation rough set. There are two essential problems to be addressed when employing the rough sets model to real-world applications as similar as the information entropy model: (1) information granulation [10,11]; (2) approximate classification realized in the presence of such induced information granules [12,13]. The idea of multi-granulation is expressed through the approximation classification realizing. For example, one of the contributions in MGRS is to describe the lower and upper approximations by the multiple equivalence relations instead of the single equivalence relation. As a matter of fact, we can construct the multi-granulation structure in the process of the information granulation. Based on this idea, the contribution of this paper includes: (1) we extend the classical information entropy model to a multi-granulation entropy model (MGE) by using a series of general binary relations; (2) moreover, a number of theorems are obtained; (3) furthermore, we employ the proposed model to evaluate the significance of the attributes for classification. A forward greedy search algorithm for feature selection is constructed. The experimental results show that the proposed method presents an effective solution for feature analysis.

The paper is organized as follows: in Section 2, some basic concepts about entropy in the view of granular computing are briefly reviewed. In Section 3, the MGE model is proposed. A series of theorems about MGE is discussed. Section 4 shows the applications of MGE to feature evaluating and feature selection. Numeric experiments are reported in Section 5. Finally, Section 6 concludes the paper.

2. Entropy in the View of Granular Computing

Knowledge representation is realized via the information system (IS) which is a tabular form, similar to databases. An information system is pair $IS = (U, A)$, where $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of objects, A is a nonempty finite set of attributes, and $f_a : U \to V_a$ is a mapping for any $a \in A$, where V_a is called the value set of a.

Relations, as a fundamental concept in mathematics, represent the connections of a set elements in the domain. A binary relation on U can be represented as a matrix. The matrix $\mathbf{M_R} = (r_{ij})_{n \times n}$ is called the relation matrix of R on the universe U. The matrix $\mathbf{M_R} = (r_{ij})_{n \times n}$ is denoted as:

$$\mathbf{M_R} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix} \tag{1}$$

In the classical set theory, the relations take values in the set $\{0,1\}$. In this case the relation matrix is a Boolean matrix. In the fuzzy set theory, the relations take values in the interval $[0,1]$. In this paper, we use a general binary relation R to denote any instantiated relation, such as fuzzy relation and kernel relation, *etc*. The fuzziness of relations is the essential characteristic in these cases. Therefore, $r_{ij} \in [0,1]$ in our study.

Given an information system $IS = (U, A)$ and a binary relation R, $U = \{x_1, x_2, \ldots, x_n\}$, $\forall x_i \in U$ the information granule $[x_i]_R$ is defined as:

$$[x_i]_R = \frac{r_{1i}}{x_1} + \frac{r_{2i}}{x_2} + \cdots + \frac{r_{ni}}{x_n} \tag{2}$$

With each sample, we express the information granule in the form of fuzzy sets. Here, we give an instance about kernel entropy to illustrate the information entropy model in the view of kernel granulation [5]:

Example 1. Given an information system $IS = (U, A)$ as follows:

Table 1. *IS* description.

	a_1	a_2
x_1	0.1	0.4
x_2	0.2	0.3

where $U = \{x_1, x_2\}$ and $A = \{a_1, a_2\}$, the two modules in the kernel entropy methodology are as follows, respectively:

(1) Information granulation:

The kernel relation is computed with Gaussian kernel as follows, where $\|x_i - x_j\|$ is the Euclidean distance between samples x_i and x_j:

$$r_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2} \right) \tag{3}$$

Hence, we have $r_{12} = r_{21} \approx 0.37$ and $r_{11} = r_{22} = 1$ if σ is set to 0.1. The kernel granules can be constructed according Equation (2).

(2) Calculating the kernel entropy:

The cardinality of $[x_i]_R$ is computed in the form of $|[x_i]_R| = \sum_{j=1}^{n} r_{ij}$. Thus, the expected cardinality of $[x_i]_R$ is computed as follows, where $|U|$ is cardinality of set U.

$$\overline{Card}([x_i]_R) = \frac{|[x_i]_R|}{|U|} \tag{4}$$

The kernel entropy is defined as follows:

$$KH(A) = -\frac{1}{|U|}\sum_{i=1}^{n}\log_2 \overline{Card}([x_i]_R) \tag{5}$$

Then, the kernel entropy of this *IS* is $KH(A) = -\frac{1}{2}\left(\log_2 \frac{1.37}{2} + \log_2 \frac{1.37}{2}\right) \approx 0.55$

Remark. To deal with nominal attributes and numerical attributes, which are common in practice, we use a extended Euclidean distance as the method introduced in literature [13]. This distance function is computed as follows:

$$\|x_i - x_j\| = \sqrt{\sum_{l=1}^{N} d_{a_l}(x_i, x_j)^2} \tag{6}$$

$$d_{a_l}(x_i, x_j) = \begin{cases} nom\_diff_{a_l}(x_i, x_j) & \text{if } a_l \text{ is a nominal attribute} \\ num\_diff_{a_l}(x_i, x_j) & \text{if } a_l \text{ is a numerical attribute} \end{cases} \tag{7}$$

where $nom\_diff_{a_l}(x_i, x_j) = \begin{cases} 0 & \text{if } x_i^l = x_j^l \\ 1 & \text{if } x_i^l \neq x_j^l \end{cases}$ and $num\_diff_{a_l}(x_i, x_j) = |x_i^l - x_j^l|$.

In a real environment, we often need to concurrently describe a target concept through multiple binary relations (e.g., neighborhood relation, kernel relation, and fuzzy relation) according to a user's requirements or targets of problem solving. Therefore, we will study the multi-granulation entropy model in the next section.

3. Multi-Granulation Entropy

In this section, two types of multi-granulation entropy (MGE) are introduced to measure the uncertainty of knowledge in information systems. Then, the joint entropy and conditional entropy are presented in the view of multi-granulation. A number of theorems will be discussed in detail.

3.1. Two Types of MGE

Definition 1. Let $IS = (U, A)$ be an information system, where $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of objects. Given a set of general binary relations $\Re = \{R_1, R_2, \ldots R_t\}$, the optimistic granule $[x_i]_{OR}$ is computed as follows, where "$\vee$" means "max":

$$[x_i]_{OR} = \overset{t}{\underset{j=1}{\vee}}[x_i]_{R_j} \tag{8}$$

The granule $[x_i]_{R_j}$ is defined in forms of fuzzy set for the "max" operation. The word "optimistic" is used to express the idea that the information granulation seeks common ground while reversing difference among these general binary relations.

Definition 2. Let $IS = (U, A)$ be an information system, where $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of objects. Given a set of general binary relations $\Re = \{R_1, R_2, \ldots R_t\}$, the pessimistic granule $[x_i]_{OR}$ is computed as follows, where "$\wedge$" means "min":

$$[x_i]_{PR} = \overset{t}{\underset{j=1}{\wedge}} [x_i]_{R_j} \tag{9}$$

The granule $[x_i]_{R_j}$ is defined in forms of fuzzy set for the "min" operation. The word "pessimistic" is used to express the idea that the information granulation seeks common ground while rejection difference among these general binary relations.

Then, the expected cardinality of $[x_i]_{OR}$ and $[x_i]_{PR}$ are computed as follows, respectively:

$$\overline{Card}([x_i]_{OR}) = \frac{|[x_i]_{OR}|}{|U|} \tag{10}$$

$$\overline{Card}([x_i]_{PR}) = \frac{|[x_i]_{PR}|}{|U|} \tag{11}$$

Here, we give the definition about the two types of MGE.

Definition 3. Let $IS = (U, A)$ be an information system, where $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of objects. Given a set of general binary relations $\Re = \{R_1, R_2, \ldots R_t\}$, $B \subseteq A$, the first type of MGE, called optimistic multi-granulation entropy (OMGE), is denoted by:

$$OH(B)_{\Re} = -\frac{1}{|U|} \sum_{i=1}^{n} \log_2 \overline{Card}([x_i]_{OR}) \tag{12}$$

Definition 4. Let $IS = (U, A)$ be an information system, where $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of objects. Given a set of general binary relations $\Re = \{R_1, R_2, \ldots R_t\}$, $B \subseteq A$, the second type of MGE, called pessimistic multi-granulation entropy (PMGE), is denoted by:

$$PH(B)_{\Re} = -\frac{1}{|U|} \sum_{i=1}^{n} \log_2 \overline{Card}([x_i]_{PR}) \tag{13}$$

The following example will illustrate the two types of MGE in detail.

Example 2. Given a nonempty finite set of objects $U = \{x_1, x_2, x_3, x_4, x_5\}$. Two relation matrixes about R_1 and R_2 are denoted as:

$$\mathbf{M}_{\mathbf{R_1}} = \begin{pmatrix} 1 & 0.4 & 0.5 & 0.6 & 0.8 \\ 0.2 & 1 & 0.4 & 0.5 & 0.1 \\ 0.6 & 0.9 & 1 & 0.1 & 0.2 \\ 0.3 & 0.5 & 0.8 & 1 & 0.2 \\ 0.4 & 0.5 & 0.6 & 0.1 & 1 \end{pmatrix}, \quad \mathbf{M}_{\mathbf{R_2}} = \begin{pmatrix} 1 & 0.6 & 0.1 & 0.2 & 0.1 \\ 0.2 & 1 & 0.8 & 0.5 & 0.1 \\ 0.6 & 0.9 & 1 & 0.2 & 0.2 \\ 0.3 & 0.7 & 0.6 & 1 & 0.3 \\ 0.4 & 0.5 & 0.8 & 0.1 & 1 \end{pmatrix} \tag{14}$$

The optimistic and pessimistic relation matrixes are denoted by $\mathbf{M_O}$ and $\mathbf{M_P}$ respectively as follows:

$$\mathbf{M_O} = \begin{pmatrix} 1 & 0.6 & 0.5 & 0.6 & 0.8 \\ 0.2 & 1 & 0.8 & 0.5 & 0.1 \\ 0.6 & 0.9 & 1 & 0.2 & 0.2 \\ 0.3 & 0.7 & 0.8 & 1 & 0.3 \\ 0.4 & 0.5 & 0.8 & 0.1 & 1 \end{pmatrix}, \quad \mathbf{M_P} = \begin{pmatrix} 1 & 0.4 & 0.1 & 0.2 & 0.1 \\ 0.2 & 1 & 0.4 & 0.5 & 0.1 \\ 0.6 & 0.9 & 1 & 0.1 & 0.2 \\ 0.3 & 0.5 & 0.6 & 1 & 0.2 \\ 0.4 & 0.5 & 0.6 & 0.1 & 1 \end{pmatrix} \tag{15}$$

Every row of the matrixes denotes the information granule (e.g., $[x_1]_{OR} = \frac{1}{x_1} + \frac{0.6}{x_2} + \frac{0.5}{x_3} + \frac{0.6}{x_4} + \frac{0.8}{x_5}$).

OMGE and PMGE are computed according to Equations (12) and (13):

$$OH(A)_{\{R_1,R_2\}} = -\frac{1}{|U|}\sum_{i=1}^{n}\log_2 \overline{Card}([x_i]_{OR}) \approx 0.54 \tag{16}$$

$$PH(A)_{\{R_1,R_2\}} = -\frac{1}{|U|}\sum_{i=1}^{n}\log_2 \overline{Card}([x_i]_{PR}) \approx 0.49 \tag{17}$$

Definition 5. Let $IS = (U, A)$ be an information system, where $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of objects. Given a set of general binary relations $\Re = \{R_1, R_2, \ldots R_t\}$, $B_1, B_2 \subseteq A$,the optimistic information granules $[x_i]_{PR}^1$ and $[x_i]_{PR}^2$ are induced by B_1 and B_2. The optimistic joint entropy is expressed as:

$$OH(B_1 \cup B_2)_{\Re} = -\frac{1}{|U|}\sum_{i=1}^{n}\log_2 \overline{Card}([x_i]_{OR}^1 \wedge [x_i]_{OR}^2) \tag{18}$$

where "$\wedge$" means "min".

The pessimistic joint entropy is defined as follows:

$$PH(B_1 \cup B_2)_{\Re} = -\frac{1}{|U|}\sum_{i=1}^{n}\log_2 \overline{Card}([x_i]_{PR}^1 \wedge [x_i]_{PR}^2) \tag{19}$$

Definition 6. Let $IS = (U, A)$ be an information system, where $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of objects. Given a set of general binary relations $\Re = \{R_1, R_2, \ldots R_t\}$, $B_1, B_2 \subseteq A$, the optimistic conditional entropy of B_2 to B_1 is expressed as:

$$OH(B_2 \mid B_1)_{\Re} = OH(B_2 \cup B_1)_{\Re} - OH(B_1)_{\Re} \tag{20}$$

Similarly, we have the pessimistic conditional entropy:

$$PH(B_2 \mid B_1)_{\Re} = PH(B_2 \cup B_1)_{\Re} - PH(B_1)_{\Re} \tag{21}$$

The conditional entropy reflects the uncertainty of B_2 if B_1 is given.

3.2. Some Theorems about MGE

Theorem 1. Let $IS = (U, A)$ be an information system, where $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of objects. Given only one equivalence binary relation $\Re = \{R_e\}$, we have:

$$OH(A)_{\Re} = PH(A)_{\Re} = H(A) \tag{22}$$

where $H(A)$ is Shannon's entropy.

Proof $OH(A)_{\Re} = PH(A)_{\Re}$ is straightforward. The equivalence binary relation is computed as:

$$r_{ij} = \begin{cases} 1 & if\ x_i = x_j \\ 0 & if\ x_i \neq x_j \end{cases} \tag{23}$$

The samples are divided into disjoint $X_1, X_2, \ldots X_m$, where $x_i = x_j \in X_k$. Assumed there are w_k samples in X_k, then $H(A) = -\sum_{i=1}^{w} \frac{w_k}{|U|}\log_2 \frac{w_k}{|U|}$. $\overline{card}([x_i]_R) = \frac{|X_k|}{|U|} = \frac{w_k}{|U|}$ if $x_i \in X_k$, we have:

$$\begin{aligned} OH(A)_{\Re} &= PH(A)_{\Re} \\ &= -\frac{1}{|U|}\sum_{i=1}^{n} \log_2 \overline{Card}([x_i]_R) \\ &= -\frac{1}{|U|}\sum_{x \in X_1} \log_2 \frac{w_1}{|U|} + \cdots + \left(-\frac{1}{|U|}\sum_{x \in X_k} \log_2 \frac{w_k}{|U|} \right) \\ &= -\frac{w_1}{|U|}\log_2 \frac{w_1}{|U|} + \cdots + \left(-\frac{w_k}{|U|}\log_2 \frac{w_k}{|U|} \right) \\ &= H(A) \end{aligned} \tag{24}$$

It is shown that the MGE is a natural generalization of the Shannon's entropy in the view of granulation by the proof above. In [14,15], the authors generalized Shannon's entropy to fuzzy entropy, kernel entropy and neighborhood entropy, respectively. These entropy models utilize solely the granularity structure of the given data, which is expressed by one suitable binary relation. The neighborhood entropy is only based on the neighborhood granulation; the fuzzy entropy on the fuzzy granulation; and the kernel entropy on the kernel granulation. Hence, it also can be concluded that the single-granulation entropy, such as neighborhood entropy, kernel entropy, fuzzy entropy, *etc.*, is the special instance of MGE.

Theorem 2. Let $IS = (U, A)$ be an information system, where $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of objects. Given a set of general binary relations $\Re = \{R_1, R_2, \ldots R_t\}$, $\Re_1 \subseteq \Re_2 \subseteq \Re$, we have:

$$OH(A)_{\Re_1} \geq OH(A)_{\Re_2} \tag{25}$$

$$PH(A)_{\Re_1} \leq PH(A)_{\Re_2} \tag{26}$$

Proof $\forall x_i \in U$, $\Re_1 \subseteq \Re_2 \subseteq \Re$, we have $[x_i]_{O\Re_1} \subseteq [x_i]_{O\Re_2}$. Therefore, $\overline{card}([x_i]_{O\Re_1}) \leq \overline{card}([x_i]_{O\Re_2})$. Obviously, $OH(A)_{\Re_1} \geq OH(A)_{\Re_2}$. Similarly, $[x_i]_{P\Re_1} \supseteq [x_i]_{P\Re_2}$, $PH(A)_{\Re_1} \leq PH(A)_{\Re_2}$.

For convenience, the monotonicity of entropy value induced by the set of relations is called the granulation monotonicity.

Corollary 1. Let $IS = (U, A)$ be an information system, where $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of objects. Given a set of general binary relations $\Re = \{R_1, R_2, \ldots R_t\}$, we have:

$$OH(A)_{\Re} \leq PH(A)_{\Re} \tag{27}$$

Proof $\forall x_i \in U$ we have $\mathop{\wedge}\limits_{j=1}^{t} [x_i]_{R_j} = [x_i]_{PR} \subseteq [x_i]_{OR} = \mathop{\vee}\limits_{j=1}^{t} [x_i]_{R_j}$. Therefore, $\overline{card}([x_i]_{PR}) \leq \overline{card}([x_i]_{OR})$. Obviously, $OH(A)_{\Re} \leq PH(A)_{\Re}$.

Corollary 2. Let $IS = (U, A)$ be an information system, where $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of objects. Given a set of general binary relations $\Re = \{R_1, R_2, \ldots R_t\}$, $\Re_1 \subseteq \Re_2 \subseteq \Re$, $B_1, B_2 \subseteq A$, we have:

$$OH(B_2 \mid B_1)_{\Re_2} \geq OH(B_2 \mid B_1)_{\Re_2} \tag{28}$$

$$PH(B_2 \mid B_1)_{\Re_1} \geq PH(B_2 \mid B_1)_{\Re_2} \tag{29}$$

Proof According to Lemma 4.1 in Ref. [16], we know that the combination of information granules by "$\vee$" operator will increase the conditional entropy monotonously. Similarly, it can be concluded that the conditional entropy will decrease through combining information granules by "$\wedge$" operator. QED

4. Feature Selection Based on MGE

One of the most important applications of information entropy theory is to evaluate the classification power of the attributes in a decision system by computing the significance of the condition attributes for the resulting decision. This entropy-based model was widely used in feature selection algorithms for categorical data [17]. However, classical entropy models cannot be used to express multi-granulation which represents the different points of view for describing one concept. Here, we show a feature selection technique based on MGE.

If the set of samples is assigned with a decision attribute D, we call this information system $IS = (U, C, D)$ a decision system, where C are conditional attributes. Therefore, as we explain in Definition 6 that multi-granulation conditional entropy $OH(D \mid C)$ ($PH(D \mid C)$) is the uncertainty of D if condition attributes C are given, conditional entropy reflects the relevance between condition attributes and decision.

Definition 7. Let $IS = (U, C, D)$ be a decision system, where $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of objects. Given a set of general binary relations $\Re = \{R_1, R_2, \ldots R_t\}$, $B \subseteq C$, we thus define significance of attribute subset B in the multi-granulation of view:

$$OSIG(B, D)_{\Re} = OH(D)_{\Re} - O(D \mid B)_{\Re} = OH(D)_{\Re} + OH(B)_{\Re} - OH(D \cup B)_{\Re} \tag{30}$$

$$PSIG(B, D)_{\Re} = PH(D)_{\Re} - P(D \mid B)_{\Re} = PH(D)_{\Re} + PH(B)_{\Re} - PH(D \cup B)_{\Re} \tag{31}$$

$OSIG(B, D)_{\Re}$ is used to evaluate the significance of attribute subset B by the optimistic multi-granulation. Similar to $OSIG(B, D)_{\Re}$, $PSIG(B, D)_{\Re}$ is another evaluation measure of the attributes. The pessimistic granules, which are formed by the binary relations $\Re$, are used to compute $PSIG(B, D)_{\Re}$. It is easy to observe that $OSIG(B, D)_{\Re}$ ($PSIG(B, D)_{\Re}$) becomes a symmetric uncertainty measure. In fact this is mutual information of B and D defined in Shannon's information theory if B and D generate Boolean equivalence relations according to Equation (23) [18]. As it is well-known, mutual information is widely applied in evaluating features and constructing decision trees [19,20], the classical definition of mutual information can just be used to deal with only one granulation. The multi-granulation significance defined here can be used to express lots of views with a series of binary relations. Equations (30) and (31) can be used to find the significant features for classification. Actually, it is impractical to get the optimal subset of features from $2^n - 1$ candidates through exhaustive search, where n is the number of features. The greedy search guided by some heuristics is usually more efficient than the plain brute-force exhaustive search. In a forward greedy search, one starts with an empty set of attributes, and keeps adding features to the subset of selected attributes one by one. Each selected attribute maximizes the increment of significance of the current subset. A forward search algorithm for feature selection based on MGE is written as follows. Here, $OSIG$ and $PSIG$ are denoted as SIG uniformly.

Algorithm 1. Feature selection based on MGE(OMGE or PMGE)

Input: decision system $IS = (U, C, D)$, binary relations $\Re = \{R_1, R_2, \ldots R_t\}$ and stopping threshold ε.

Output: selected features red.

```
$red \leftarrow \phi$
while $red \neq C$
  for each $a_i \in (C - red)$
      compute $sig_i = SIG(a_i \cup red, D)$
  end for
  find the maximal $sig_i$ and the corresponding attribute $a_i$
  if $sig_i - SIG(red, D) > \varepsilon$
      $red \leftarrow red \cup a_i$
  else
      exit while
  end if
end while
return $red$
```

The time complexity of the algorithm is $O(n^2 m \log m)$, where n and m are the numbers of features and samples, respectively. It is worth noting that the proposed measures of mutual information can be incorporated with other search strategies used in other feature selection algorithms, such as ABB (Automatic Branch and Bound), probabilistic search [21] and GP (Genetic programming) [22]. In this study, we are not going to compare the influence of search strategies on the results of feature selection. Here we focus on the comparison of the proposed method when dealing with different evaluation measures.

5. Experimental Analysis

In this section, we compare the effectiveness of MGE in evaluating feature quality. The data sets are downloaded from the UCI Machine Learning Repository. They are described in Table 2. The numerical attributes of the samples are linearly normalized as follows:

$$x = (x - x_{\min})/(x_{\max} - x_{\min}) \tag{32}$$

where $x_{\min}$ and $x_{\max}$ are the bounds of the given attribute. Three popular leaning algorithms such as CART, liner SVM and RBF SVM are introduced to evaluate the quality of selected features. The experiments were run in a 10-fold cross validation mode. The parameters of the linear SVM and RBF SVM are taken as the default values (the use of the MATLAB toolkit osu_svm3.00).

Table 2. Data description.

ID	Data	Samples	Features	Class
1	wine	178	13	3
2	wdbc	569	31	2
3	iono	351	34	2
4	heart	270	13	2
5	glass	214	9	7
6	wpbc	198	33	2
7	sonar	208	60	2

In the experiment, we employ three symmetric membership functions for multi-granulation. One is the kernel relation defined as Equation (3) in Example 1; the other two are computed as follows, respectively:

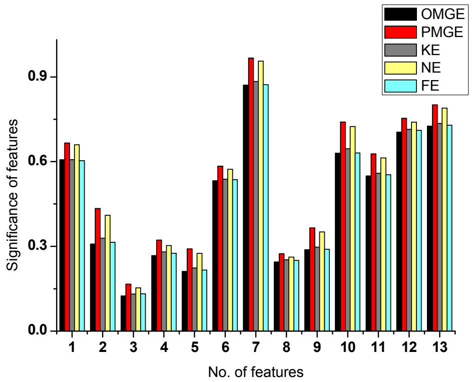
$$r_{ij} = \begin{cases} 1 \ \ if \left\| x_i - x_j \right\| \le \delta \\ 0 \ \ if \left\| x_i - x_j \right\| > \delta \end{cases} \tag{33}$$

$$r_{ij} = \begin{cases} 1 - 4 \times \left\| x_i - x_j \right\| \ \ , if \left\| x_i - x_j \right\| \le 0.25 \\ 0 \ \ , if \left\| x_i - x_j \right\| > 0.25 \end{cases} \tag{34}$$

Equation (33), called neighborhood relation, is used to compute neighborhood entropy (NE) in Ref. [6] where the threshold $\delta \ge 0$. According to this definition, the samples in a neighborhood granule have the distance is less than the threshold δ. Literature [6] has explained that the result is optimum if threshold δ is set between 0.1 and 0.2. In the following, if not specified, $\delta = 0.15$. Similarly, the fuzzy entropy (FE) is proposed based on the fuzzy relation according to Equation (34) [4]. We compare MGE with kernel entropy (KE), NE and FE, where the compared methods are the typical single-granulation entropy. The parameters of the KE and NE are kept consistent in Ref. [5] and Ref. [6]. We compute the significance of single feature with five evaluation functions, such as OMGE, PMGE, KE, NE and FE. At the same time, we reported the classification accuracies of the each feature based on the use of the linear SVM and RBF SVM.

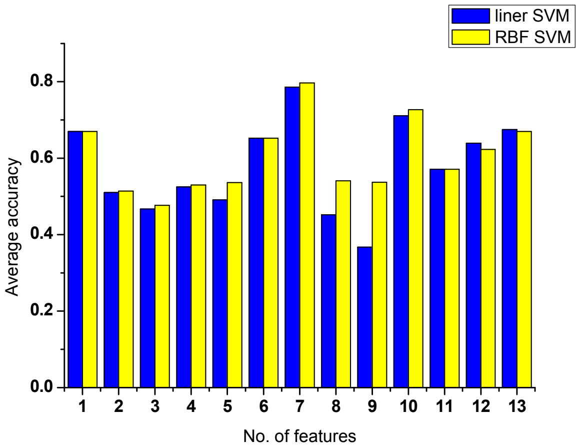
Two data sets wine and glass are used in the experiment. There are 13 features in the wine and nine features in the glass dataset. The results are given in Figures 1 and 2. As to the wine data, the features 1, 6, 7, 10, 11, 12, 13 produce higher values of all evaluation functions, as shown in Figure 1a; at the same time, we can also find that the classification accuracies of these features are better than others (again shown in Figure 1b). As to the glass data, features 2, 3, 4, 8 are better than others in terms of the five evaluating functions, corresponding the classification accuracies of features 2, 3, 4, 8 are also higher than the other features. These results show that all five evaluating functions can produce good estimates of classification ability of the features. It can be concluded that OMGE and PMGE are competent with other entropy models.

Figure 1. Significance and accuracy of single feature (**wine**). (**a**) Significance of a single feature computed with different evaluating. (**b**) Classification accuracies obtained for single features when using linear SVM and RBF SVM.



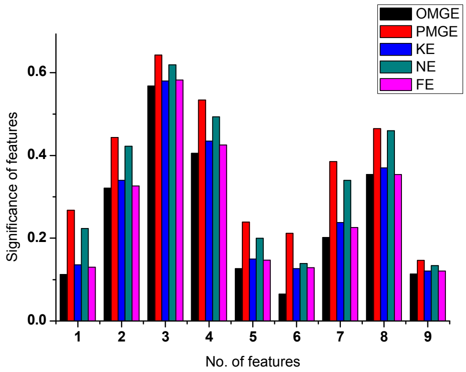
(a)

Figure 1. *Cont.*

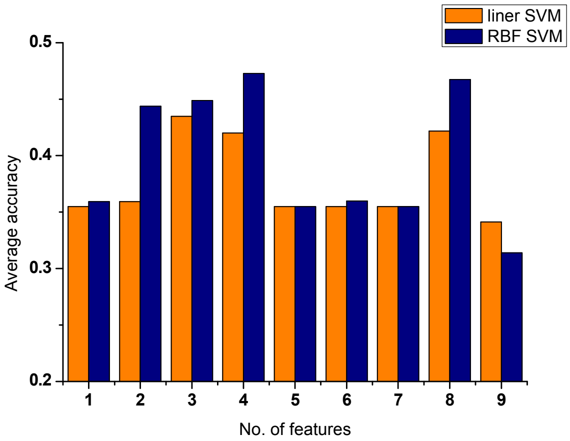


(b)

Figure 2. Significance and accuracy of single feature(**glass**). (**a**) Significance of a single feature computed with different evaluating. (**b**) Classification accuracies obtained for single features when using linear SVM and RBF SVM.



(a)



(b)

The above results show MGE can be used to evaluate single attributes. Now, we show the effectiveness in attribute reduction. The selected features with different algorithms are presented in Tables 3 and 4, respectively. Regarding OMGE, PMGE, FE, NE and KE, the orders of the features presented in the tables are the orders that the features are kept being added to the feature space. These orders reflect the relative significance of features in terms of the corresponding measures. Some results can be derived from the selected attributes. First, whatever attribute selection techniques have been used, most of the attributes in all datasets can be deleted. The reduction rate is high to 90% for some datasets, such as sonar and wpbc. Second, some selected attributes are slightly different. Especially, some of the selected features are the subset of attributes selected by other models.

Table 3. Subsets of features selected with OMGE and PMGE.

Data	OMGE	PMGE
wine	7,1,10,13	7,1,11,4
wdbc	29,22,23,12,9	24,29,23,30,26,9,13,10,28,3,27
iono	5,6,8,25,28,24,10,21	5,6,34,29,8,23
heart	13,12,3,11,1,7,4	13,12,3,1,10,4
glass	3,7,4,9,5	3,7,4,9,5,1
wpbc	34,2,13,14,7	2,34,13,7,23
sonar	12,27,21,37,32,30,54	12,16,26,40,48

Table 4. Subsets of features selected with FE, NE and KE.

Data	FE	NE	KE
wine	7, 1, 10,13	7,1,11,4	7,1,10,13
wdbc	29,22,23,12,9	24,29,23,30,8,27,26,13,10,3,19	29,22,23,9,12
iono	5,6,8,25,28,24,34,7	5,6,34,29,8,23	5,6,8,25,28,24,34,7
heart	13,12,3,10,1,7,11,2,8,4	13,12,3,10,1,4,5	13,12,3,10,1,7,11
glass	3,7,4,9,5	3,7,4,9,5,1	3,7,4,9,5
wpbc	34,2,13,14,7	2,34,13,7,23	34,2,13,14,7
sonar	12,27,21,37,32,30,54	12,16,26,40,48	12,16,26,37,22,32,28

As we know, we consider the ranking of features in feature selection, sometimes, a little difference in feature qualities may lead to completely different ranking. Therefore, the great difference between these selected features is the difference between the qualities of features computed with diverse granularities. In other words, there is a inconsistent relationship between its values under one-granularity and those under the another granularity. In [7], the authors give a tentative study that multi-granulation model will display its advantage for rule extraction when two granularities process a contradiction relationship. We will test this idea by the following experiment. We build classification models with the selected features and test their classification performance based on 10-fold cross validation. The average value and standard deviation are used to measure the classification performance. We compare the raw data, MGE, FE, NE and KE in Tables 5–7, where learning algorithms CART, linear SVM and RBF SVM are introduced to evaluate the selected features.

Table 5. Classification accuracies based on CART (%).

Data	Raw data	OMGE	PMGE	FE	NE	KE
wine	86.4 ± 7.9	92.2 ± 7.5	89.9 ± 8.5	92.2 ± 7.5	89.9 ± 8.5	92.2 ± 7.5
wdbc	90.3 ± 6.0	93.0 ± 3.8	93.5 ± 3.9	93.0 ± 3.8	94.0 ± 3.2	93.0 ± 3.8
iono	86.4 ± 7.2	87.5 ± 5.6	88.6 ± 6.5	88.1 ± 6.0	88.6 ± 6.5	88.1 ± 6.0
heart	77.0 ± 5.5	78.5 ± 7.3	80.4 ± 9.0	75.2 ± 9.2	80.0 ± 7.7	77.8 ± 9.2
glass	69.2 ± 13.2	65.7 ± 12.9	65.1 ± 14.6	65.7 ± 12.9	65.1 ± 14.6	65.7 ± 12.9
wpbc	70.2 ± 5.4	70.7 ± 10.3	71.1 ± 11.9	70.7 ± 10.3	71.1 ± 11.9	70.7 ± 10.3
sonar	57.7 ± 9.2	73.1 ± 12.6	72.2 ± 15.3	73.1 ± 12.6	72.2 ± 15.3	62.6 ± 13.2

Table 6. Classification accuracies based on liner SVM (%).

Data	Raw data	OMGE	PMGE	FE	NE	KE
wine	98.3 ± 2.7	97.2 ± 3.9	94.4 ± 5.2	97.2 ± 3.9	94.4 ± 5.2	97.2 ± 3.9
wdbc	98.0 ± 1.9	96.1 ± 2.1	96.3 ± 2.1	96.1 ± 2.1	95.9 ± 2.1	96.1 ± 2.1
iono	87.5 ± 6.4	83.4 ± 5.3	85.0 ± 5.9	85.0 ± 5.3	85.0 ± 5.9	85.0 ± 5.3
heart	84.1 ± 9.3	83.0 ± 8.9	81.9 ± 7.2	82.9 ± 9.4	82.2 ± 6.0	82.6 ± 8.6
glass	55.7 ± 7.7	60.4 ± 9.1	57.1 ± 8.7	60.4 ± 9.1	57.1 ± 8.7	60.4 ± 9.1
wpbc	77.3 ± 5.7	76.3 ± 3.0	76.3 ± 3.0	76.3 ± 3.0	76.3 ± 3.0	76.3 ± 3.0
sonar	64.4 ± 15.7	64.9 ± 11.6	67.8 ± 15.7	64.9 ± 11.6	67.8 ± 15.7	65.5 ± 13.5

Table 7. Classification accuracies based on RBF SVM (%).

Data	Raw data	OMGE	PMGE	FE	NE	KE
wine	97.8 ± 2.9	96.7 ± 3.9	95.0 ± 4.1	96.7 ± 3.9	95.0 ± 4.1	96.7 ± 3.9
wdbc	97.0 ± 2.6	96.7 ± 2.1	97.2 ± 2.3	96.7 ± 2.1	97.5 ± 2.5	96.7 ± 2.1
iono	94.0 ± 4.2	93.7 ± 4.8	92.6 ± 5.7	94.0 ± 4.6	92.6 ± 5.7	94.0 ± 4.6
heart	79.2 ± 9.6	81.1 ± 7.9	83.3 ± 7.6	82.6 ± 9.4	84.0 ± 7.8	81.5 ± 9.2
glass	68.3 ± 12.1	62.7 ± 11.9	64.1 ± 11.2	62.7 ± 11.9	64.1 ± 11.2	62.7 ± 11.9
wpbc	78.9 ± 6.0	74.3 ± 7.4	75.3 ± 7.7	74.3 ± 7.4	75.3 ± 7.7	74.3 ± 7.4
sonar	58.5 ± 16.0	65.9 ± 14.9	64.1 ± 11.2	65.9 ± 14.9	64.1 ± 11.2	61.1 ± 7.2

Comparing the performance of raw data and granulation-based selection, we can find although most of features have been removed, most of the classification accuracies derived from the reduced data sets do not decrease, but increase. It shows there are redundant and irrelevant attributes in the raw data.

The experimental results show that no matter which classification algorithms are used, MGE is better than or equivalent to KE. Table 6 shows that MGE outperforms FE and NE with respect to liner SVM. As to CART learning algorithm in Table 5, MGE is better than or equivalent to NE for six of the seven databases. It can be concluded that MGE is a better choice for the diverse granularities. Actually, the different decision makers have different granulation points of view. Therefore, it is necessary to take diverse factors into consideration for granular computing in the real world.

6. Conclusions

In this paper, the classical single-granulation entropy theory has been extended. As a result of this extension, a multi-granulation entropy model (MGE) has been developed. The uncertainty of the information system is defined by using multiple relations on the universe. These relations can be chosen according to a user's requirements or targets of problem solving.

In MGE model, we introduce OMGE and PMGE to describe the relations between different granularities. Based on the mutual information defined through MGE, we proposed the forward greed features selection algorithms, which will be helpful for applying this theory to practical issues. MGE provides an effective approach in the context of multiple granulations. We conclude that the single-granulation entropy is the special instance of MGE. The experimental result shows that MGE will display its advantage for rule extraction and knowledge discovery when the different granularities in information systems possess a contradiction or inconsistent relationship.

The future work could move along two directions. First, the existing feature selection algorithms based entropy sometimes might not be robust enough for real-world applications. How to improve it is an important issue. Second, we will continue to construct MGE models with various binary relations for discussing the common properties of this kind of entropy model.

Acknowledgments

This work is supported by the Fundation of Science & Technology Department of Sichuan Province, China (Grant No. 2012GZ0061).

Conflict of Interest

The authors declare no conflict of interest.

References and Notes

1. Knuth, K.H. Lattice duality: The origin of probability and entropy. *Neurocomputing* **2005**, *67*, 245–274.
2. Harremoës, P.; Vajda, I. On the bahadur-efficient testing of uniformity by means of the entropy. *IEEE Trans. Inform. Theory* **2008**, *54*, 321–331.
3. Santhanam, N.P.; Modha, D. Lossy Lempel-Ziv like compression algorithms for memoryless sources. In Proceedings of 49th Annual Allerton Conference, Monticello, IL, USA, 28–30 September 2011; pp. 1751–1756.
4. Yu, D.R.; Hu, Q.H.; Wu, C.X. Uncertainty measures for fuzzy relations and their applications. *Appl. Soft Comput.* **2007**, *7*, 1135–1143.
5. Hu, Q.H.; Zhang, L.; Chen, D.G.; Pedrycz, W.; Yu, D.R. Gaussian kernel based fuzzy rough sets: model, uncertainty measures and applications. *Int. J. Approx. Reason.* **2010**, *51*, 453–471.
6. Hu, Q.H.; Zhang, L.; Zhang, D.; Pan, W.; An, S.; Pedrycz, W. Measuring relevance between discrete and continuous features based on neighborhood mutual information. *Expert Syst. Appl.* **2011**, *38*, 10737–10750.

7. Qian, Y.H.; Liang, J.Y.; Yao, Y.Y.; Dang, C.Y. MGRS: A multi-granulation rough set. *Inform. Sci.* **2010**, *180*, 949–970.
8. Lin, G.P.; Li, J.J. A covering-based pessimistic multigranulation rough set. In proceedings of International Conference on Intelligent Computing, Zhengzhou, China, 11–14 August, 2011; pp. 673–680.
9. Xu, W.H.; Zhang, X.T.; Wang, Q.R. A generalized multi-granulation rough set approach. In proceedings of International Conference on Intelligent Computing, Zhengzhou, China, 11–14 August, 2011; pp. 681–689.
10. Qin, K.Y.; Yang, J.L.; Pei, Z. Generalized rough sets based on reflexive and transitive relations. *Inform. Sci.* **2008**, *178*, 4138–4141.
11. Zhu, W.; Wang, S.P. Rough matroids based on relations. *Inform. Sci.* **2013**, *232*, 241–252.
12. Tang, J.G.; She, K.; Min, F.; Zhu, W. A matroidal approach to rough set theory. *Theor. Comput. Sci.* **2013**, *471*, 1–11.
13. Jing, S.Y.; She, K.; Ali, S. A Universal neighbourhood rough sets model for knowledge discovering from incomplete heterogeneous data. *Expert Syst.* **2013**, *30*, 89–96.
14. Al-Sharhan, S.; Karray, F.; Gueaieb, W.; Basir, O. Fuzzy entropy: a brief survey. In *The 10th IEEE International Conference on Fuzzy Systems*; IEEE: Melbourne, Australia, 2001; Volume 3, pp. 1135–1139.
15. Hu, Q.H.; Yu, D.R. Neighborhood Entropy. In Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, China, 12–15 July 2009; Volume 3, pp. 1776–1782.
16. Wang, G.Y. Rough reduction in algebra view and information view. *Int. J. Intell. Syst.* **2003**, *18*, 679–688.
17. Slezak, D. Approximate entropy reducts. *Fund. Informat.* **2002**, *53*, 365–390.
18. Hu, Q.H.; Yu, D.R.; Xie, Z.X.; Liu, J.F. Fuzzy probabilistic approximation spaces and their information measures. *IEEE Trans. Fuzzy Syst.* **2006**, *14*, 191–201.
19. Yu, L.; Liu, H. Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res.* **2004**, *5*, 1205–1224.
20. Battiti, R. Using mutual information for selecting features in supervised neural net learning. *IEEE T. Neural Networ.* **1994**, *5*, 537–550.
21. Dash, M.; Liu, H. Consistency-based search in feature selection. *Artif. Intell.* **2003**, *151*, 155–176.
22. Muni, D.P.; Pal, N.R.; Das, J. Genetic programming for simultaneous feature selection and classifier design. *IEEE T. Syst. Man. Cy. B* **2006**, *36*, 106–117.