


Article

# A Fast Feature Selection Algorithm by Accelerating Computation of Fuzzy Rough Set-Based Information Entropy

Xiao Zhang <sup>1,\*</sup> , Xia Liu <sup>1</sup> and Yanyan Yang <sup>2</sup>

<sup>1</sup> Department of Applied Mathematics, School of Sciences, Xi'an University of Technology, Xi'an 710048, China; liuxia@xaut.edu.cn

<sup>2</sup> Department of Automation, Tsinghua University, Beijing 100084, China; yangyanyan@mail.tsinghua.edu.cn

\* Correspondence: zhangxiao@xaut.edu.cn; Tel.: +86-029-82066369

Received: 30 August 2018; Accepted: 9 October 2018; Published: 13 October 2018



**Abstract:** The information entropy developed by Shannon is an effective measure of uncertainty in data, and the rough set theory is a useful tool of computer applications to deal with vagueness and uncertainty data circumstances. At present, the information entropy has been extensively applied in the rough set theory, and different information entropy models have also been proposed in rough sets. In this paper, based on the existing feature selection method by using a fuzzy rough set-based information entropy, a corresponding fast algorithm is provided to achieve efficient implementation, in which the fuzzy rough set-based information entropy taking as the evaluation measure for selecting features is computed by an improved mechanism with lower complexity. The essence of the acceleration algorithm is to use iterative reduced instances to compute the lambda-conditional entropy. Numerical experiments are further conducted to show the performance of the proposed fast algorithm, and the results demonstrate that the algorithm acquires the same feature subset to its original counterpart, but with significantly less time.

**Keywords:** information entropy; fuzzy rough set theory; feature selection; fast algorithm

## 1. Introduction

Rough set theory [1] presented by Pawlak in 1982 is a useful tool to deal with vagueness and uncertainty information in the field of computer sciences. The research of rough set theory has mainly focused on both the generalizations of rough set models and the applications in different data environments, which has already attached much attention in granular computing [2–4], feature selection [5–8], dynamic data mining [9–11], and big data mining [12,13]. On the other hand, since the information entropy is powerful to measure information uncertainty, it has been extensively applied in practical problems, such as decision making [14], time series [15], portfolio selection [16], and so on.

In view of the effectiveness of information entropy to measure uncertainty in formation, information entropy has been extensively applied in the rough set theory to mine knowledge, which mainly concentrates on constructing rough set-based entropy in different information systems to measure the significance of features (or attributes) or the quality of knowledge granules and on exploring practical applications of rough set-based entropy. Specifically, in the aspect of constructing rough set-based entropy [17–28], the references [18] and [19] respectively introduced the concepts of information entropy, rough entropy, and knowledge granulation in complete and incomplete information systems and provided their important properties. Hu et al. [20] proposed the generalizations of the entropy to calculate the information of a fuzzy approximation space and a fuzzy probabilistic approximation space, respectively. Xu et al. [21] introduced the definition of rough entropy of rough

sets in ordered information systems. Mi et al. [22] formulated the entropy of the generalized fuzzy approximation space. Dai and Tian [25] provided the concepts of knowledge information entropy and knowledge rough entropy in set-valued information systems, and investigated their properties. Dai et al. [26] presented the rough decision entropy to evaluate the uncertainty of interval-valued decision systems. Chen et al. [27] introduced the neighborhood entropy to evaluate the uncertainty of neighborhood information systems. Wang et al. [28] put forward a unified form of uncertainty measures for general binary relations.

In the aspect of exploring practical applications of rough set-based entropy [29–35], Pal et al. [31] defined the measure “rough entropy of image” for image object extraction in the framework of rough sets. Tsai et al. [32] provided an entropy-based fuzzy rough classification approach to acquire classification rules. Chen and Wang [33] presented an improved clustering algorithm based on both rough set theory and entropy theory. Sen and Pal [34] gave classes of entropy measures based on rough set theory to quantify the grayness and spatial ambiguity in images. Chen et al. [35] put forward an entropy-based gene selection method based on the neighborhood rough set model. Furthermore, it is worth noting that one of the most important applications of rough set-based entropy is feature selection (attribute reduction) [36–44]. For example, Miao and Hu [36] defined the significance of attributes from the viewpoint of information and then proposed a heuristic attribute reduction algorithm by using the mutual information. Wang et al. [37] developed two novel heuristic attribute reduction algorithms based on the conditional information entropy. Hu et al. [39] introduced a fuzzy entropy to measure the uncertainty in kernel approximation based on fuzzy rough sets, and thus proposed the feature evaluation index and a feature selection algorithm. Sun et al. [40] provided the rough entropy-based uncertainty measures for feature selection in incomplete decision systems. Liang et al. [41] introduced the incremental mechanisms for three representative information entropies and then developed a group incremental entropy-based feature selection algorithm based on the rough set theory with multiple instances being added to a decision system. Chen et al. [43] proposed a neighborhood entropy to select feature subset based on the neighborhood rough set model. Zhang et al. [44] presented a feature selection method by using the fuzzy rough set-based information entropy.

Since the computation of the fuzzy rough set-based information entropy in [44] is quite time-consuming, we propose in this paper a corresponding improved mechanism with lower complexity to compute the entropy and develop a fast feature selection algorithm that can quickly obtain the same result to the feature selection algorithm in [44]. In addition, the performance of the fast algorithm is shown by some numerical experiment.

In the remainder of this paper, we briefly review in Section 2 the feature selection algorithm in [44] and some related knowledge. In Section 3, the computational properties of the fuzzy rough set-based information entropy in [44] are presented. A fast feature selection approach with lower complexity has been developed. Numerical experiments were documented in Section 4 to show the performance of the proposed fast feature selection algorithm.

## 2. Preliminaries

As indicated in [45], a fuzzy information system is a pair  $(U, A)$  in which  $U = \{x_1, x_2, \dots, x_n\}$  is the universe of discourse and  $A = \{a_1, a_2, \dots, a_m\}$  is the attribute set. For each attribute  $a_t \in A$ , a mapping  $a_t : U \rightarrow V_{a_t}$  holds where  $V_{a_t}$  is the domain of  $a_t$ , and a fuzzy relation  $R_{\{a_t\}}$  can be defined. The fuzzy relation of a subset  $B \subseteq A$  is  $R_B = \bigcap_{a_t \in B} R_{\{a_t\}}$ .

It is possible to define the corresponding fuzzy relations for the attributes with different types of values, and one can refer to [44] for the details. Here, a fuzzy relation  $R$  is a fuzzy set that is defined on the fuzzy power set  $F(U \times U)$  to measure the similarity between two objects in the universe  $U$ .

By adding an attribute set  $D = \{d\}$  with  $A \cap D = \emptyset$  into a fuzzy information system  $(U, A)$ , we obtain a fuzzy decision system  $(U, A \cup D)$  where  $A$  is the conditional attribute set and  $D$  is the

decision attribute set. It should be pointed out that  $d$  is a nominal attribute on which a mapping  $d : U \rightarrow V_d$  holds and  $V_d$  is the domain of  $d$ .

By utilizing a fuzzy rough sets-based information entropy, a forward addition feature selection algorithm is proposed in [44], and it is as follows.

---

**Algorithm 1:** Computing an  $\varepsilon$ -approximate reduct of a fuzzy decision system.

---

**Input** : A fuzzy decision system  $(U, A \cup D)$  with  $U = \{x_1, x_2, \dots, x_n\}$ , and a parameter  $\varepsilon \geq 0$ .

**Output:** An  $\varepsilon$ -approximate reduct  $B$ .

```

1 Initialize  $B = \emptyset, H_\lambda(D|B) = n/e$ ;
2 for  $i = 1$  to  $n$  do
3   | compute  $\lambda_i = \underline{R}_A[x_i]_D(x_i)$ ;
4 end
5 while  $H_\lambda(D|B) \geq \varepsilon$  do
6   | for each  $a_i \in A \setminus B$  do
7     | compute  $SIG_\lambda(a_i, B, D) = H_\lambda(D|B) - H_\lambda(D|B \cup \{a_i\})$ ;
8   end
9   | choose an attribute  $a_{i_0}$  satisfying  $SIG_\lambda(a_{i_0}, B, D) = \max_i SIG_\lambda(a_i, B, D)$ ;
10  | let  $H_\lambda(D|B) = H_\lambda(D|B \cup \{a_{i_0}\})$ ;
11  | if  $H_\lambda(D|B) \geq \varepsilon$  then
12    |  $B = B \cup \{a_{i_0}\}$ ;
13  end
14 end
15 return  $B$ ;

```

---

In Step 3 of Algorithm 1,  $\underline{R}_A[x_i]_D$  is the fuzzy lower approximation of the decision class  $[x_i]_D$  based on the fuzzy relation  $R_A$ , which is proposed in the pioneering work of fuzzy approximation operators [46] and is concretely computed by

$$\underline{R}_A[x_i]_D(x_j) = \inf_{x_j \in U} \max\{1 - R_A(x_i, x_j), [x_i]_D(x_j)\}. \tag{1}$$

Here,  $[x_i]_D$  is the crisp decision class to which the object  $x_i$  belongs, and  $[x_i]_D = \{x_j \in U : (x_i, x_j) \in R_D\}$  where  $R_D$  is the equivalence relation generated by the nominal decision attribute  $d$ . Thus, the membership function of the decision class  $[x_i]_D$  is

$$[x_i]_D(x_j) = \begin{cases} 1, & x_j \in [x_i]_D; \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

In Step 7,  $SIG_\lambda(a_i, B, D)$  is the significance of the attribute  $a_i$  ( $a_i \in A \setminus B$ ) for  $B$  relative to  $D$ , which is factually the decrease of the  $\lambda$ -conditional entropy in the process of adding one attribute. Here, the  $\lambda$ -conditional entropy of the decision attribute set  $D$  relative to the conditional attribute subset  $B$ , i.e.,  $H_\lambda(D|B)$ , is defined in [44] as

$$H_\lambda(D|B) = -\frac{1}{n} \sum_{i=1}^n \left( |[x_i]_B^{\lambda_i} \cap [x_i]_D| \log \frac{|[x_i]_B^{\lambda_i} \cap [x_i]_D|}{|[x_i]_B^{\lambda_i}|} \right) \tag{3}$$

where

$$[x_i]_B^{\lambda_i}(x_j) = \begin{cases} \lambda_i, & 1 - R_B(x_i, x_j) < \lambda_i; \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

is the fuzzy granule of  $x_i$  with respect to  $B$ , and  $\lambda_i = \underline{R}_A[x_i]_D(x_i)$ .

It should be pointed out that  $|X|$  is the cardinality of the fuzzy set  $X$ , which is defined in [38] as  $|X| = \sum_{i=1}^n X(x_i)$ . For example,  $|[x_i]_B^{\lambda_i}| = \sum_{i=1}^n [x_i]_B^{\lambda_i}(x_i)$ . Moreover, as indicated in [44], if there exists an object  $x_{i_0} \in U$  such that  $\lambda_{i_0} = \underline{R}_A[x_{i_0}]_D(x_{i_0}) = 0$ , then take

$$|[x_{i_0}]_B^0 \cap [x_{i_0}]_D| \log \frac{|[x_{i_0}]_B^0 \cap [x_{i_0}]_D|}{|[x_{i_0}]_B^0|} = 0. \tag{5}$$

Generally, the  $\lambda$ -conditional entropy is less than  $n/e$ . Thus, the  $\lambda$ -conditional entropy  $H_\lambda(D|B)$  is initialized to  $n/e$  in Step 1 of Algorithm 1. Furthermore, the  $\lambda$ -conditional entropy is of monotonicity, i.e.,  $H_\lambda(D|C) \geq H_\lambda(D|B)$  holds for  $C \subseteq B \subseteq A$ , which yields  $SIG_\lambda(a_i, B, D) \geq 0$ . Therefore, in the iteration procedure of Algorithm 1, the feature  $a_{i_0}$  satisfying  $SIG_\lambda(a_{i_0}, B, D) = \max_i SIG_\lambda(a_i, B, D)$  is added in a feature subset.

As indicated in [44], the time complexity of Algorithm 1 is  $O(|U|^2|A|^2)$ , in which Step 7 is the critical step to select features and the complexity of computing  $SIG_\lambda(a_i, B, D)$  is  $O(|U|^2)$ , as well as the complexity of running Steps 2–4 is  $O(|U|^2|A|)$ . Here,  $|\cdot|$  is the cardinality of one crisp set. Computing  $SIG_\lambda(a_i, B, D)$  may require a great amount of time if  $|U|$  is large. Therefore, a natural idea of accelerating Algorithm 1 is that accelerating the computation of  $SIG_\lambda(a_i, B, D)$  according to computational properties of the  $\lambda$ -conditional entropy.

### 3. Accelerated Computation of $\lambda$ -Conditional Entropy

In the following, we concentrate on the computational characteristic of  $\lambda$ -conditional entropy. Firstly, we review the following theorem in [44].

**Theorem 1.** Let  $(U, A)$  be a fuzzy information system with a fuzzy relation  $R_B$  for each  $B \subseteq A$ . For any fuzzy set  $X \in F(U)$ ,

$$\underline{R}_B X(x_i) = \sup\{\lambda : [x_i]_B^\lambda \subseteq X\}. \tag{6}$$

Here,  $[x_i]_B^\lambda$  with  $\lambda \leq \underline{R}_B X(x_i)$  is a basic fuzzy granule with respect to  $B$  to characterize the inner structure of  $X$ . Let  $X$  be  $[x_i]_D$ . Then,  $[x_i]_B^{\lambda_i}$  with  $\lambda_i = \underline{R}_B[x_i]_D(x_i)$  is the biggest granule contained in  $[x_i]_D$ .

Let  $(U, A \cup D)$  be a fuzzy decision system with  $U = \{x_1, x_2, \dots, x_n\}$  and  $B \subseteq A$ . Denote

$$U_B^* = \left\{ x_i : |[x_i]_B^{\lambda_i} \cap [x_i]_D| < |[x_i]_B^{\lambda_i}|, x_i \in U \right\} \tag{7}$$

as the object set in which each object  $x_i$  satisfies  $|[x_i]_B^{\lambda_i} \cap [x_i]_D| < |[x_i]_B^{\lambda_i}|$ . It is obvious to have  $U_B^* \subseteq U$ . We then have the following property.

**Property 1.** Let  $(U, A \cup D)$  be a fuzzy decision system with  $U = \{x_1, x_2, \dots, x_n\}$  and  $B \subseteq A$ . Then

$$H_\lambda(D|B \cup \{a\}) = -\frac{1}{|U|} \sum_{x_i \in U_B^*} \left( |[x_i]_{B \cup \{a\}}^{\lambda_i} \cap [x_i]_D| \log \frac{|[x_i]_{B \cup \{a\}}^{\lambda_i} \cap [x_i]_D|}{|[x_i]_{B \cup \{a\}}^{\lambda_i}|} \right) \tag{8}$$

holds for any  $a \in A \setminus B$ .

**Proof.** Assume that  $U \setminus U_B^* \neq \emptyset$ . Then, for any  $x_i \in U \setminus U_B^*$ , we have  $|[x_i]_B^{\lambda_{i_0}} \cap [x_i]_D| = |[x_i]_B^{\lambda_i}|$ , which is equivalent to  $[x_i]_B^{\lambda_i} \subseteq [x_i]_D$ . Then, according to Theorem 1, it is obtained that  $\underline{R}_B[x_i]_D(x_i) \geq \lambda_i$ .

Because of  $R_B[x_i]_D(x_i) \leq R_A[x_i]_D(x_i) = \lambda_i$ , we have  $R_B[x_i]_D(x_i) = \lambda_i$ , which yields  $R_{B \cup \{a\}}[x_i]_D(x_i) = \lambda_i$  and then  $[x_i]_{B \cup \{a\}}^{\lambda_i} \subseteq [x_i]_D$  for any  $a \in A \setminus B$ . Therefore,  $[x_i]_{B \cup \{a\}}^{\lambda_i} \cap [x_i]_D = [x_i]_{B \cup \{a\}}^{\lambda_i}$  and then

$$-\frac{1}{|U|} \left| [x_i]_{B \cup \{a\}}^{\lambda_i} \cap [x_i]_D \right| \log \frac{\left| [x_i]_{B \cup \{a\}}^{\lambda_i} \cap [x_i]_D \right|}{\left| [x_i]_{B \cup \{a\}}^{\lambda_i} \right|} = 0,$$

which yields

$$\begin{aligned} H_\lambda(D|B \cup \{a\}) &= -\frac{1}{|U|} \sum_{x_i \in U} \left( \left| [x_i]_{B \cup \{a\}}^{\lambda_i} \cap [x_i]_D \right| \log \frac{\left| [x_i]_{B \cup \{a\}}^{\lambda_i} \cap [x_i]_D \right|}{\left| [x_i]_{B \cup \{a\}}^{\lambda_i} \right|} \right) \\ &= -\frac{1}{|U|} \sum_{x_i \in U_B^*} \left( \left| [x_i]_{B \cup \{a\}}^{\lambda_i} \cap [x_i]_D \right| \log \frac{\left| [x_i]_{B \cup \{a\}}^{\lambda_i} \cap [x_i]_D \right|}{\left| [x_i]_{B \cup \{a\}}^{\lambda_i} \right|} \right) \\ &\quad - \frac{1}{|U|} \sum_{x_i \in U \setminus U_B^*} \left( \left| [x_i]_{B \cup \{a\}}^{\lambda_i} \cap [x_i]_D \right| \log \frac{\left| [x_i]_{B \cup \{a\}}^{\lambda_i} \cap [x_i]_D \right|}{\left| [x_i]_{B \cup \{a\}}^{\lambda_i} \right|} \right) \\ &= -\frac{1}{|U|} \sum_{x_i \in U_B^*} \left( \left| [x_i]_{B \cup \{a\}}^{\lambda_i} \cap [x_i]_D \right| \log \frac{\left| [x_i]_{B \cup \{a\}}^{\lambda_i} \cap [x_i]_D \right|}{\left| [x_i]_{B \cup \{a\}}^{\lambda_i} \right|} \right) \end{aligned}$$

□

Assume that the similarity relation  $R_B(x_i, x_j)$  has been computed for any  $x_i \in U$  and  $x_j \in U$ . Then, according to Property 1, the time complexity of  $H_\lambda(D|B \cup \{a\})$  is  $O(|U_B^*||U|)$ , which is generally less than  $O(|U|^2)$  since  $U_B^* \subseteq U$  holds.

Denote

$$U_B^{x_i} = \{x_j : [x_i]_B^{\lambda_i}(x_j) = \lambda_i, x_j \in U\} \tag{9}$$

as the object set in which each object belongs to the fuzzy set  $[x_i]_B^{\lambda_i}$  with the degree being  $\lambda_i$ . Since

$$[x_i]_B^{\lambda_i}(x_j) = \begin{cases} \lambda_i, & 1 - R_B(x_i, x_j) < \lambda_i; \\ 0, & \text{otherwise,} \end{cases}$$

then, for any  $x_j \in U \setminus U_B^{x_i}$ , it is easily obtained that  $[x_i]_B^{\lambda_i}(x_j) = 0$ . Furthermore, we have the following property.

**Property 2.** Let  $(U, A \cup D)$  be a fuzzy decision system with  $U = \{x_1, x_2, \dots, x_n\}$  and  $B \subseteq A$ . Then, for any  $a \in A \setminus B$ , we have

$$\left| [x_i]_{B \cup \{a\}}^{\lambda_i} \right| = \sum_{x_j \in U_B^{x_i}} [x_i]_{B \cup \{a\}}^{\lambda_i}(x_j) \tag{10}$$

and

$$\left| [x_i]_{B \cup \{a\}}^{\lambda_i} \cap [x_i]_D \right| = \sum_{x_j \in (U_B^{x_i} \cap [x_i]_D)} [x_i]_{B \cup \{a\}}^{\lambda_i}(x_j). \tag{11}$$

**Proof.** Assume that  $U \setminus U_B^{x_i} \neq \emptyset$ . Then, for any  $a \in A \setminus B$  and any  $x_j \in U \setminus U_B^{x_i}$ , it is obtained that the fuzzy similarity relation  $R_{B \cup \{a\}} = R_B \cap R_{\{a\}} \subseteq R_B$  and  $1 - R_B(x_i, x_j) \geq \lambda_i$ , which yields  $1 - R_{B \cup \{a\}}(x_i, x_j) \geq 1 - R_B(x_i, x_j) \geq \lambda_i$  and then  $[x_i]_{B \cup \{a\}}^{\lambda_i}(x_j) = 0$ . Therefore, we have

$$\begin{aligned}
 |[x_i]_{B \cup \{a\}}^{\lambda_i}| &= \sum_{x_j \in U} [x_i]_{B \cup \{a\}}^{\lambda_i}(x_j) \\
 &= \sum_{x_j \in U_B^{x_i}} [x_i]_{B \cup \{a\}}^{\lambda_i}(x_j) + \sum_{x_j \in U \setminus U_B^{x_i}} [x_i]_{B \cup \{a\}}^{\lambda_i}(x_j) \\
 &= \sum_{x_j \in U_B^{x_i}} [x_i]_{B \cup \{a\}}^{\lambda_i}(x_j)
 \end{aligned}$$

and

$$\begin{aligned}
 |[x_i]_{B \cup \{a\}}^{\lambda_i} \cap [x_i]_D| &= \sum_{x_j \in [x_i]_D} [x_i]_{B \cup \{a\}}^{\lambda_i}(x_j) \\
 &= \sum_{x_j \in (U_B^{x_i} \cap [x_i]_D)} [x_i]_{B \cup \{a\}}^{\lambda_i}(x_j) + \sum_{x_j \in ((U \setminus U_B^{x_i}) \cap [x_i]_D)} [x_i]_{B \cup \{a\}}^{\lambda_i}(x_j) \\
 &= \sum_{x_j \in (U_B^{x_i} \cap [x_i]_D)} [x_i]_{B \cup \{a\}}^{\lambda_i}(x_j).
 \end{aligned}$$

□

Substituting Equations (10) and (11) into Equation (8), we then have

$$H_\lambda(D|B \cup \{a\}) = -\frac{1}{|U|} \sum_{x_i \in U_B^*} \left( \left( \sum_{x_j \in (U_B^{x_i} \cap [x_i]_D)} [x_i]_{B \cup \{a\}}^{\lambda_i}(x_j) \right) \log \frac{\left( \sum_{x_j \in (U_B^{x_i} \cap [x_i]_D)} [x_i]_{B \cup \{a\}}^{\lambda_i}(x_j) \right)}{\left( \sum_{x_j \in U_B^{x_i}} [x_i]_{B \cup \{a\}}^{\lambda_i}(x_j) \right)} \right). \tag{12}$$

**Corollary 1.** Let  $(U, A \cup D)$  be a fuzzy decision system with  $U = \{x_1, x_2, \dots, x_n\}$  and  $B \subseteq A$ . Then, for any  $a \in A \setminus B$ , we have

$$U_{B \cup \{a\}}^{x_i} \subseteq U_B^{x_i}. \tag{13}$$

**Proof.** For any  $x_j \in U \setminus U_B^{x_i}$ , we have  $[x_i]_B^{\lambda_i}(x_j) = 0$ . It can be obtained from the proof process of Property 2 that  $[x_i]_{B \cup \{a\}}^{\lambda_i}(x_j) = 0$  holds for any  $a \in A \setminus B$ , which yields  $x_j \in U \setminus U_{B \cup \{a\}}^{x_i}$ . Thus,  $(U \setminus U_B^{x_i}) \subseteq (U \setminus U_{B \cup \{a\}}^{x_i})$ , which implies  $U_{B \cup \{a\}}^{x_i} \subseteq U_B^{x_i}$ . □

Assume that the similarity relation  $R_B(x_i, x_j)$  has been computed for any  $x_i \in U$  and  $x_j \in U$ . Then, according to Equation (12), the time complexity of  $H_\lambda(D|B \cup \{a\})$  is  $O(C|U_B^*|)$ , which is generally less than  $O(|U|^2)$  since both  $C \leq |U|$  and  $|U_B^*| \leq |U|$  hold. Here,  $C = \max\{|U_B^{x_i}| : x_i \in U_B^*\}$ . Therefore, according to Properties 1 and 2, we can use Equation (12) to compute  $H_\lambda(D|B \cup \{a\})$  and then obtain an accelerated algorithm in the following.

Compared with Algorithm 1, there exist three aspects of differences in Algorithm 2. First, Algorithm 2 needs to set  $U_B^*$  and  $U_B^{x_i}$  ( $x_i \in U$ ) to  $U$  in Steps 1–4. Second, the evaluation measure  $H_\lambda(D|B \cup \{a_i\})$  is improved to compute according to Equation (12) in Step 10, in which  $U_{B \cup \{a_i\}}^*$  can be automatically acquired without additional computation. Here, the complexity of computing  $H_\lambda(D|B \cup \{a_i\})$  is  $O(C|U_B^*|)$ , where  $C = \max\{|U_B^{x_i}| : x_i \in U_B^*\}$ . Third,  $U_B^*$  and  $U_B^{x_i}$  ( $x_i \in U$ ) are iteratively updated in Steps 16–20, and Steps 17–20 need  $O(C|U_B^*|)$ . Furthermore, the main procedure of Algorithm 2 for selecting features, namely Steps 8–22, needs to be run at most  $|A|$  times, so the time complexity is  $O(C|U_B^*||A|^2)$ . However, the main process Steps 5–14 in Algorithm 1 for selecting features requires  $O(|U|^2|A|^2)$ . It should be pointed out that both  $|U_B^*|$  and  $C$  may monotonously decrease in the iteration process of Algorithm 2, which mainly contributes to accelerate computation.

---

**Algorithm 2:** Accelerating computation of an  $\varepsilon$ -approximate reduct of a fuzzy decision system.

---

**Input** : A fuzzy decision system  $(U, A \cup D)$  with  $U = \{x_1, x_2, \dots, x_n\}$ , and a parameter  $\varepsilon \geq 0$ .

**Output:** An  $\varepsilon$ -approximate reduct  $B$ .

```

1 Initialize  $B = \emptyset$ ,  $H_\lambda(D|B) = n/e$ , and  $U_B^* = U$ ;
2 for  $i = 1$  to  $n$  do
3   | initialize  $U_B^{x_i} = U$ ;
4 end
5 for  $i = 1$  to  $n$  do
6   | compute  $\lambda_i = \underline{R}_A[x_i]_D(x_i)$ ;
7 end
8 while  $H_\lambda(D|B) \geq \varepsilon$  do
9   | for each  $a_i \in A \setminus B$  do
10    | compute  $H_\lambda(D|B \cup \{a_i\})$  according to Equation (12) (%Note:  $U_{B \cup \{a_i\}}^*$  can be obtained
11    | in the computation);
12   | end
13   | choose an attribute  $a_{i_0}$  satisfying  $H_\lambda(D|B \cup \{a_{i_0}\}) = \min_i H_\lambda(D|B \cup \{a_i\})$ ;
14   | let  $H_\lambda(D|B) = H_\lambda(D|B \cup \{a_{i_0}\})$ ;
15   | if  $H_\lambda(D|B) \geq \varepsilon$  then
16     |  $B = B \cup \{a_{i_0}\}$ ;
17     |  $U_B^* = U_{B \cup \{a_{i_0}\}}^*$ ;
18     | for each  $x_i \in U_B^*$  do
19       | compute  $U_{B \cup \{a_{i_0}\}}^{x_i} = \{x_j : [x_i]_B^{\lambda_i}(x_j) = \lambda_i, x_j \in U_B^{x_i}\}$  (%Note: The equation is due to
20       | Equation (16));
21       |  $U_B^{x_i} = U_{B \cup \{a_{i_0}\}}^{x_i}$ ;
22     | end
23   | end
24 end
25 return  $B$ ;

```

---

#### 4. Numerical Experiment

In this section, numerical experiments are conducted to assess the performance of Algorithm 2. The experiment mainly focuses on showing the computational efficiency of Algorithm 2. In order to achieve the task, nine data sets are downloaded from UCI Repository of machine learning databases. The data sets are briefly described in Table 1.

**Table 1.** Description of the data sets.

Data Set	Abbreviation of Data Set	Number of Objects	Number of Conditional Attributes			Number of Classes
			All	Nominal	Real-Valued	
Horse Colic	Horse	368	22	15	7	2
Credit Approval	Credit	690	15	9	6	2
German Credit Data	German	1000	20	13	7	2
Wisconsin Diagnostic Breast Cancer	WDBC	569	30	0	30	2
Libras Movement	Libras	360	90	0	90	15
Musk (Version 1)	Musk1	476	166	0	166	2
Hill-Valley	HV	606	100	0	100	2
Wall-Following Robot Navigation Data	Robot	5456	24	0	24	4
Waveform Database Generator (Version 2)	WDG2	5000	40	0	40	3

#### 4.1. Pretreatment of the Data Sets and Design of the Experiment

For each data set, the object set, conditional attribute set and decision attribute set are denoted by  $U$ ,  $A$ , and  $D$ , respectively. If there are some real-valued conditional attributes in  $A$ , then, for each real-valued attribute  $a \in A$ , the attribute value of each object is normalized according to the method in [44] as

$$\tilde{a}(x_i) = \frac{a(x_i) - \min_j a(x_j)}{\max_j a(x_j) - \min_j a(x_j)}, \quad x_i \in U, \quad (14)$$

so that  $\tilde{a}(x_i) \in [0, 1]$  for each  $x_i \in U$ . Here,  $a$  is still used to denote the corresponding normalized conditional attribute for notational simplicity.

The experiment was designed as follows. Given one of the pretreated data sets, the objects were randomly divided into 20 approximately equal parts. The first part was taken as the 1st data set, the combination of both the first and the second parts was regarded as the 2nd data set, the combination of the anterior three parts was regarded as the 3rd data set, ..., and the combination of all twenty parts was taken as the 20th data set. For each of the generated 20 data sets, a fuzzy relation for each normalized conditional attribute  $a$  is defined as

$$R_{\{a\}}(x_i, x_j) = 1 - |a(x_i) - a(x_j)|, \quad x_i, x_j \in U_k. \quad (15)$$

On the other hand, a special fuzzy relation, namely an equivalence relation, is defined for each nominal attribute  $a \in A$  by

$$R_{\{a\}}(x_i, x_j) = \begin{cases} 1, & a(x_i) = a(x_j); \\ 0, & \text{otherwise} \end{cases}, \quad (16)$$

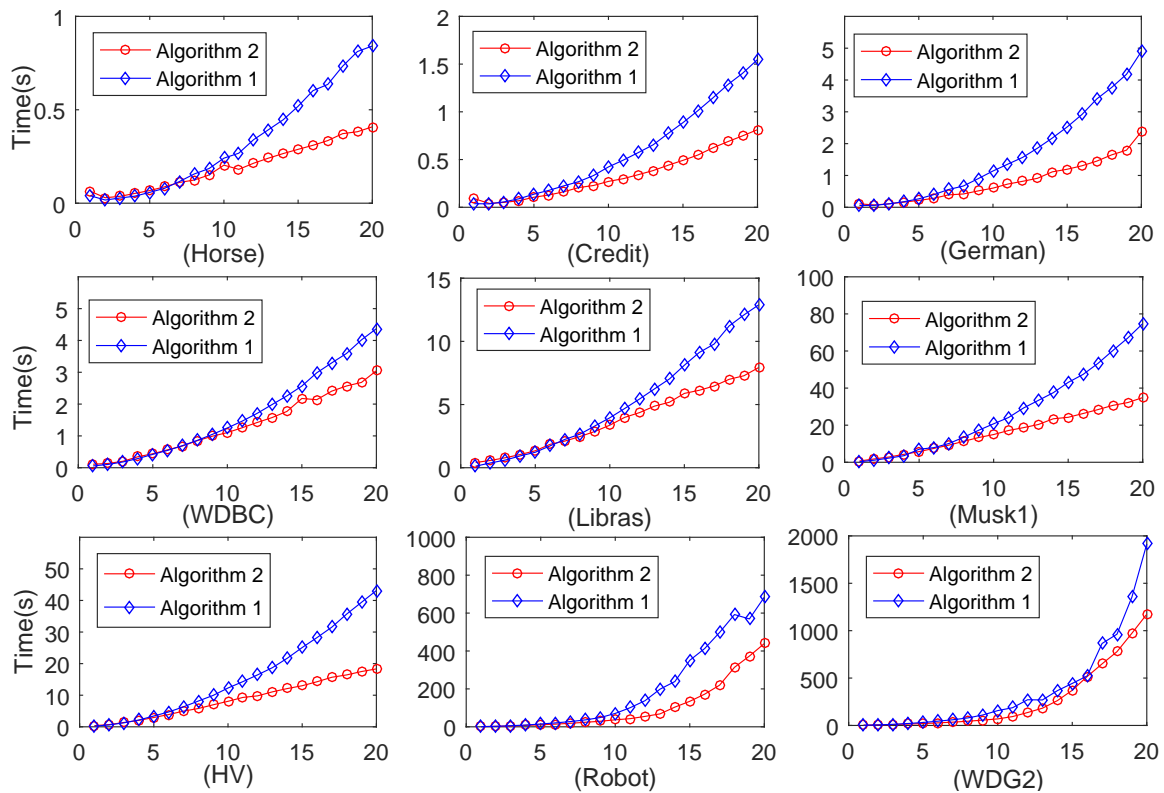
where  $x_i, x_j \in U_k$ . Here,  $U_k$  is the universe determined by the  $k$ -th data set. In this way, a fuzzy decision system  $(U_k, A \cup D)$  is formed for the  $k$ -th data set. Then, Algorithms 1 and 2 were used to obtain the computation time of these fuzzy decision systems, respectively. Furthermore, the "ten-fold approach" was also used to access the efficiency of the fast algorithm proposed in this paper. Specifically, for each of the pretreated data sets, the instances were randomly divided into 10 approximately equal parts. The  $k$ -th part was removed and the remainder was taken as the  $k$ -th data set, which generates the ten data sets called the ten-fold data sets. Then, the fuzzy relations for real-valued attributes and nominal attributes were defined according to Equations (18) and (19), respectively, which then formed a fuzzy decision system for each of the ten-fold data sets. Algorithms 1 and 2 were used to obtain the computation time of the fuzzy decision systems, respectively. Moreover, it should be pointed out that the output results obtained by both Algorithms 1 and 2 are the same for the same threshold values  $\varepsilon$ . The parameter  $\varepsilon$  determines the number of the selected features. The smaller the threshold value  $\varepsilon$  is, the more selected features there are and thus the more computation time is needed. Therefore, the parameter  $\varepsilon$  in both Algorithms 1 and 2 was set to 0. The experiment was performed by MATLAB R2016a on a personal computer with Intel(R) Core(TM) i7-4510U CPU @2.00 GHz configuration, 8 G Memory, and the 64-bit Windows 7 system.

#### 4.2. Comparison of Computation Time of Algorithms 1 and 2

##### 4.2.1. Comparison of Computation Time on 20 Data Sets Generated by Each Data Set

The computation time on 20 data sets generated by each data set respectively obtained by Algorithms 1 and 2 is depicted in Figure 1. For each of the sub-figures in Figure 1, the x-coordinate indicates the generated data sets and the number  $k$  expresses the  $k$ -th data set. In other words, the x-coordinate expresses the size of each data set and the number  $k$  is factually  $(5 * k)\%$  data of original data sets. On the other hand, the y-coordinate shows the running time (in seconds).





**Figure 1.** Computation time of Algorithms 1 and 2 with the increase of the size of each data set.

It is seen from Figure 1 that, for each data set, with the increase in data size, both Algorithms 1 and 2 require more time. At the beginning, the two algorithms cost an almost equivalent amount of time. Algorithm 2 needs a little more time relative to Algorithm 1 since the advantage of Algorithm 2 is limited by a smaller data set size. Algorithm 2 may need more time to run Steps 17–20. However, with the increase in data set size, Algorithm 2 obviously requires less running time than Algorithm 1. Therefore, the proposed Algorithm 2 is efficient and can be regarded as an accelerated version of Algorithm 1.

#### 4.2.2. Comparison of Computation Time on Ten-Folds Data Sets Produced by Each Data Set

The computation time of ten-fold data sets generated by each data set is depicted in Figure 2. For each of the sub-figures in Figure 2, the x-coordinate indicates the generated data sets and the number  $i$  expresses the  $i$ -th data set, and the y-coordinate shows the running time (in second). Furthermore, the average computation time is listed in Table 2. In addition, the average cardinalities of the selected feature subset, which is expressed by  $|\cdot|$ , are also listed in the 3rd and 5th columns of Table 2. Moreover, in order to illustrate the variation tendency of  $|U_B^*|$  in the iteration process of the proposed Algorithm 2, the relevant result obtained by one of the ten-fold data sets is depicted in Figure 3. For each of the sub-figures in Figure 3, the x-coordinate indicates the number of iterations in Algorithm 2 and the y-coordinate expresses the cardinality of  $U_B^*$ .

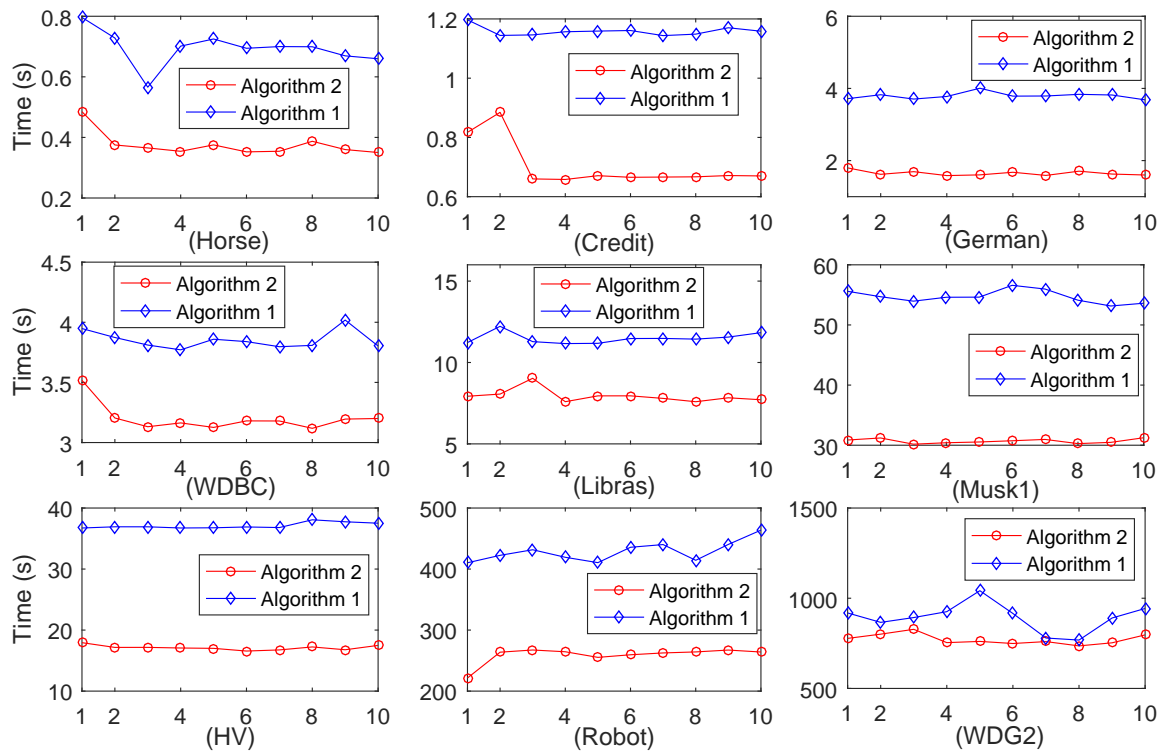
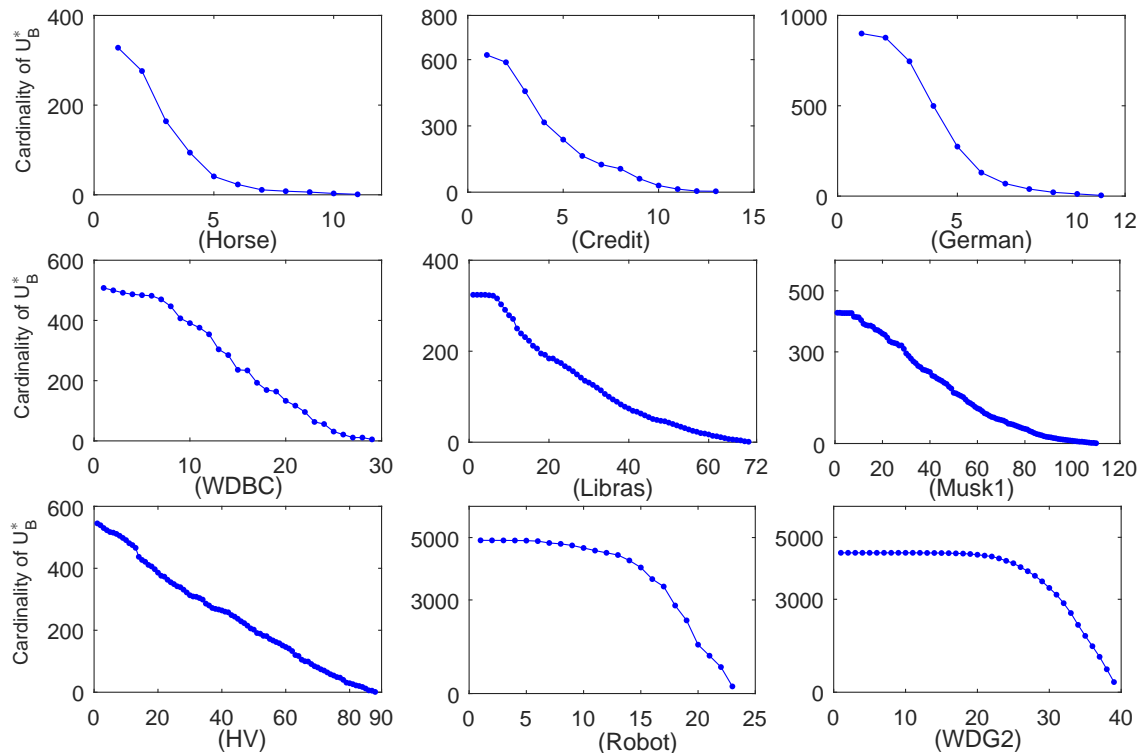


Figure 2. Computation time of Algorithms 1 and 2 on ten-fold data sets generated by each data set.

Table 2. Average results of Algorithms 1 and 2 obtained from the ten-fold data sets.

Data Set	Algorithm 2		Algorithm 1 [44]	
	Average Running Time (s)	·	Average Running Time (s)	·
Horse	0.38	12.7	0.69	12.7
Credit	0.70	13.9	1.16	13.9
German	1.65	12.9	3.79	12.9
WDBC	3.20	30.0	3.85	30.0
Libras	7.94	71.4	11.48	71.4
Musk1	30.69	112.4	54.69	112.4
HV	17.12	90.0	37.11	90.0
Robot	259.00	24.0	428.85	24.0
WDG2	771.46	40.0	894.15	40.0

It can be clearly seen in Figure 2 and Table 2 that, for each of the data sets, Algorithm 2 requires less time than Algorithm 1 for the ten-fold data sets. Especially for data sets German, Musk1, HV, and Robot, Algorithm 2 requires much less time and needs approximately no greater than 60% of the running time of Algorithm 1. Thus, it seems that Algorithm 2 requires significantly less running time for the data sets with a larger size or with more features. Moreover, the results of the 3rd and the 5th columns in Table 2 verify that the selected features respectively obtained by Algorithms 1 and 2 are the same. In addition, it can be seen from Figure 3 that  $|U_B^*|$  does monotonously decrease with the increase of the iteration number. In fact, the decrease of  $|U_B^*|$  contributes to the accelerating computation of Algorithm 2. Therefore, Algorithm 2 is validated to be effective again on the ten-fold data sets.



**Figure 3.** Variation of  $|U_B^*|$  with the increase of iteration number in Algorithm 2.

## 5. Conclusions

Based on the existing feature selection algorithm, by utilizing a fuzzy rough set-based information entropy, an accelerated feature selection algorithm according to the computational properties of fuzzy rough set-based information entropy, in which the entropy is computed by a lower time complexity, is presented in this paper. The numerical experiment results demonstrate that the algorithm can effectively decrease computation time and thus is efficient and effective. In future work, the proposed fast feature selection algorithm will be considered to deal with a dynamic data environment in which new instances or new features are added.

**Author Contributions:** “Conceptualization, X.Z.; Data curation, X.Z. and Y.Y.; Formal analysis, X.L.; Funding acquisition, X.Z.; Methodology, X.Z.; Project administration, X.Z., X.L. and Y.Y.; Software, X.Z. and Y.Y.; Validation, X.L.; Writing—Original draft, X.Z.” All authors have read and approved the final manuscript.

**Acknowledgments:** The authors thank the reviewers for their valuable comments and suggestions. This work was supported by the National Natural Science Foundation of China (Nos. 61602372, 61806162 and 61806108), the PhD Research Startup Foundation of Xi’an University of Technology (No. 109-256081504) and the China Postdoctoral Science Foundation (No. 2018M631475).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356. [[CrossRef](#)]
2. Lin, T.Y.; Yao, Y.Y.; Zadeh, L.A. *Data Mining, Rough Sets and Granular Computing*; Physica-Verlag: Heidelberg, Germany, 2002.
3. Qian, Y.H.; Zhang, H.; Sang, Y.L.; Liang, J.Y. Multigranulation decision-theoretic rough sets. *Int. J. Approx. Reason.* **2014**, *55*, 225–237. [[CrossRef](#)]
4. Luo, C.; Li, T.R.; Yi, Z.; Fujita, H. Matrix approach to decision-theoretic rough sets for evolving data. *Knowl.-Based Syst.* **2016**, *99*, 123–134. [[CrossRef](#)]
5. Zhao, S.Y.; Tsang, E.; Chen, D.G. The model of fuzzy variable precision rough sets. *IEEE Trans. Fuzzy Syst.* **2009**, *17*, 451–467. [[CrossRef](#)]

6. Qian, Y.H.; Liang, J.Y.; Pedrycz, W.; Dang, C.Y. Positive approximation: An accelerator for attribute reduction in rough set theory. *Artif. Intell.* **2010**, *174*, 597–618. [[CrossRef](#)]
7. Wang, C.Z.; Qi, Y.L.; Shao, M.W.; Hu, Q.H.; Chen, D.G.; Qian, Y.H.; Lin, Y.J. A fitting model for feature selection with fuzzy rough sets. *IEEE Trans. Fuzzy Syst.* **2017**, *25*, 741–753. [[CrossRef](#)]
8. Zhang, X.; Mei, C.L.; Chen, D.G.; Yang, Y.Y. A fuzzy rough set-based feature selection method using representative instances. *Knowl.-Based Syst.* **2018**, *151*, 216–229. [[CrossRef](#)]
9. Ananthanarayana, V.; Murty, M.N.; Subramanian, D. Tree structure for efficient data mining using rough sets. *Pattern Recognit. Lett.* **2003**, *24*, 851–862. [[CrossRef](#)]
10. Chen, H.M.; Li, T.R.; Luo, C.; Horng, S.J.; Wang, G.Y. A decision-theoretic rough set approach for dynamic data mining. *IEEE Trans. Fuzzy Syst.* **2015**, *23*, 1958–1970. [[CrossRef](#)]
11. Yang, Y.Y.; Chen, D.G.; Wang, H. Active sample selection based incremental algorithm for attribute reduction with rough sets. *IEEE Trans. Fuzzy Syst.* **2017**, *25*, 825–838. [[CrossRef](#)]
12. Hu, Q.H.; Zhang, L.J.; Zhou, Y.C.; Pedrycz, W. Large-scale multimodality attribute reduction with multi-kernel fuzzy rough sets. *IEEE Trans. Fuzzy Syst.* **2018**, *26*, 226–238. [[CrossRef](#)]
13. Qian, J.; Xia, M.; Yue, X.D. Parallel knowledge acquisition algorithms for big data using MapReduce. *Int. J. Mach. Learn. Cybern.* **2018**, *9*, 1007–1021. [[CrossRef](#)]
14. Ye, J.; Cui, W.H. Exponential entropy for simplified neutrosophic sets and its application in decision making. *Entropy* **2018**, *20*, 357. [[CrossRef](#)]
15. Girault, J.M.; Humeau-Heurtier, A. Centered and averaged fuzzy entropy to improve fuzzy entropy precision. *Entropy* **2018**, *20*, 287. [[CrossRef](#)]
16. Zhou, R.X.; Liu, X.; Yu, M.; Huang, K. Properties of risk measures of generalized entropy in portfolio selection. *Entropy* **2017**, *19*, 657. [[CrossRef](#)]
17. Düntsch, I.; Gediga, G. Uncertainty measures of rough set prediction. *Artif. Intell.* **1998**, *106*, 109–137. [[CrossRef](#)]
18. Liang, J.Y.; Shi, Z.Z. The information entropy, rough entropy and knowledge granulation in rough set theory. *Int. J. Uncertain. Fuzz. Knowl.-Based Syst.* **2004**, *12*, 37–46. [[CrossRef](#)]
19. Liang, J.Y.; Shi, Z.Z.; Li, D.Y.; Wierman, M.J. Information entropy, rough entropy and knowledge granulation in incomplete information systems. *Int. J. Gen. Syst.* **2006**, *35*, 641–654. [[CrossRef](#)]
20. 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. [[CrossRef](#)]
21. Xu, W.H.; Zhang, X.Y.; Zhang, W.X. Knowledge granulation, knowledge entropy and knowledge uncertainty measure in ordered information systems. *Appl. Soft Comput.* **2009**, *9*, 1244–1251. [[CrossRef](#)]
22. Mi, J.S.; Leung, Y.; Zhao, H.Y.; Feng, T. Generalized fuzzy rough sets determined by a triangular norm. *Inf. Sci.* **2008**, *178*, 3203–3213. [[CrossRef](#)]
23. Qian, Y.H.; Liang, J.Y. Combination entropy and combination granulation in rough set theory. *Int. J. Uncertain. Fuzz. Knowl.-Based Syst.* **2008**, *16*, 179–193. [[CrossRef](#)]
24. Ma, W.M.; Sun, B.Z. Probabilistic rough set over two universes and rough entropy. *Int. J. Approx. Reason.* **2012**, *53*, 608–619. [[CrossRef](#)]
25. Dai, J.H.; Tian, H.W. Entropy measures and granularity measures for set-valued information systems. *Inf. Sci.* **2013**, *240*, 72–82. [[CrossRef](#)]
26. Dai, J.H.; Wang, W.T.; Xu, Q.; Tian, H.W. Uncertainty measurement for interval-valued decision systems based on extended conditional entropy. *Knowl.-Based Syst.* **2012**, *27*, 443–450. [[CrossRef](#)]
27. Chen, Y.M.; Wu, K.S.; Chen, X.H.; Tang, C.H.; Zhu, Q.X. An entropy-based uncertainty measurement approach in neighborhood systems. *Inf. Sci.* **2014**, *279*, 239–250. [[CrossRef](#)]
28. Wang, C.Z.; He, Q.; Shao, M.W.; Xu, Y.Y.; Hu, Q.H. A unified information measure for general binary relations. *Knowl.-Based Syst.* **2017**, *135*, 18–28. [[CrossRef](#)]
29. Beaubouef, T.; Petry, F.E.; Arora, G. Information-theoretic measures of uncertainty for rough sets and rough relational databases. *Inf. Sci.* **1998**, *109*, 185–195. [[CrossRef](#)]
30. Jiang, F.; Sui, Y.F.; Cao, C.G. An information entropy-based approach to outlier detection in rough sets. *Expert Syst. Appl.* **2010**, *37*, 6338–6344. [[CrossRef](#)]
31. Pal, S.K.; Shankar, B.U.; Mitra, P. Granular computing, rough entropy and object extraction. *Pattern Recognit. Lett.* **2005**, *26*, 2509–2517. [[CrossRef](#)]
32. Tsai, Y.C.; Cheng, C.H.; Chang, J.R. Entropy-based fuzzy rough classification approach for extracting classification rules. *Expert Syst. Appl.* **2006**, *31*, 436–443. [[CrossRef](#)]

33. Chen, C.B.; Wang, L.Y. Rough set-based clustering with refinement using Shannon's entropy theory. *Comput. Math. Appl.* **2006**, *52*, 1563–1576. [[CrossRef](#)]
34. Sen, D.; Pal, S.K. Generalized rough sets, entropy, and image ambiguity measures. *IEEE Trans. Syst. Man Cybern. B* **2009**, *39*, 117–128. [[CrossRef](#)] [[PubMed](#)]
35. Chen, Y.; Zhang, Z.; Zheng, J.; Ma, Y.; Xue, Y. Gene selection for tumor classification using neighborhood rough sets and entropy measures. *J. Biomed. Inform.* **2017**, *67*, 59–68. [[CrossRef](#)] [[PubMed](#)]
36. Miao, D.Q.; Hu, G.R. A heuristic algorithm for reduction of knowledge. *J. Comput. Res. Dev.* **1999**, *36*, 681–684.
37. Wang, G.Y.; Yu, H.; Yang, D.C. Decision table reduction based on conditional information entropy. *Chin. J. Comput.* **2002**, *25*, 759–766.
38. Hu, Q.H.; Yu, D.R.; Xie, Z.X. Information-preserving hybrid data reduction based on fuzzy-rough techniques. *Pattern Recognit. Lett.* **2006**, *27*, 414–423. [[CrossRef](#)]
39. 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. [[CrossRef](#)]
40. Sun, L.; Xu, J.C.; Tian, Y. Feature selection using rough entropy-based uncertainty measures in incomplete decision systems. *Knowl.-Based Syst.* **2012**, *36*, 206–216. [[CrossRef](#)]
41. Liang, J.Y.; Wang, F.; Dang, C.Y.; Qian, Y.H. A group incremental approach to feature selection applying rough set technique. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 294–308. [[CrossRef](#)]
42. Foithong, S.; Pinngern, O.; Attachoo, B. Feature subset selection wrapper based on mutual information and rough sets. *Expert Syst. Appl.* **2012**, *39*, 574–584. [[CrossRef](#)]
43. Chen, Y.M.; Xue, Y.; Ma, Y.; Xu, F.F. Measures of uncertainty for neighborhood rough sets. *Knowl.-Based Syst.* **2017**, *120*, 226–235. [[CrossRef](#)]
44. Zhang, X.; Mei, C.L.; Chen, D.G.; Li, J.H. Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy. *Pattern Recognit.* **2016**, *56*, 1–15. [[CrossRef](#)]
45. Chen, D.G. *Theory and Methods of Fuzzy Rough Sets*; Science Press: Beijing, China, 2013.
46. Dubois, D.; Prade, H. Rough fuzzy sets and fuzzy rough sets. *Int. J. Gen. Syst.* **1990**, *17*, 191–209. [[CrossRef](#)]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).