


Article

# Neighborhood Attribute Reduction: A Multicriterion Strategy Based on Sample Selection

Yuan Gao <sup>1</sup>, Xiangjian Chen <sup>1,\*</sup>, Xibei Yang <sup>1,\*</sup> and Pingxin Wang <sup>2</sup>

<sup>1</sup> School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China; maxgaoyuan@gmail.com

<sup>2</sup> School of Science, Jiangsu University of Science and Technology, Zhenjiang 212003, China; wangpingxin@just.edu.cn

\* Correspondence: chenxiangjian66@gmail.com (X.C.); jsjxy\_yxb@just.edu.cn (X.Y.); Tel.: +86-180-1280-9731 (X.C.); +86-177-5136-0234 (X.Y.)

Received: 16 September 2018; Accepted: 10 November 2018; Published: 16 November 2018



**Abstract:** In the rough-set field, the objective of attribute reduction is to regulate the variations of measures by reducing redundant data attributes. However, most of the previous concepts of attribute reductions were designed by one and only one measure, which indicates that the obtained reduct may fail to meet the constraints given by other measures. In addition, the widely used heuristic algorithm for computing a reduct requires to scan all samples in data, and then time consumption may be too high to be accepted if the size of the data is too large. To alleviate these problems, a framework of attribute reduction based on multiple criteria with sample selection is proposed in this paper. Firstly, cluster centroids are derived from data, and then samples that are far away from the cluster centroids can be selected. This step completes the process of sample selection for reducing data size. Secondly, multiple criteria-based attribute reduction was designed, and the heuristic algorithm was used over the selected samples for computing reduct in terms of multiple criteria. Finally, the experimental results over 12 UCI datasets show that the reducts obtained by our framework not only satisfy the constraints given by multiple criteria, but also provide better classification performance and less time consumption.

**Keywords:** attribute reduction; cluster centroid; multiple criteria; rough set; sample selection

## 1. Introduction

Rough sets [1,2], firstly proposed by Pawlak, have been demonstrated to be useful in data mining [3,4], artificial intelligence [5], decision analysis [6,7], and so on. As one of the important strategies of feature selection, attribute reduction in rough-set theory plays a key role, since it provides us with clear semantic explanations of the selected attributes. Those semantic explanations can be reflected by constraints in terms of the considered measures, such as approximation quality and conditional entropy. For example, Hu et al. [8] have studied uncertainty measures related to fuzzy rough sets, and then further explored approximation quality-based attribute reduction; Dai et al. [9–11] investigated attribute reduction with respect to several types conditional entropies; Wang et al. [12] not only proposed a conditional discrimination index for overcoming the limitations of conditional entropies, but also provided the corresponding approach to attribute reduction.

It must be noted that most of the previous results about attribute reduction are based on the consideration of a single measure. For example, if attribute reduction is designed to preserve the approximation quality, then intuitively it may not perform well in learning tasks. This is mainly because that approximation quality is only a measure of uncertainty, which is slightly related to, for example,

classification, clustering, or regression. From this point of view, multiple criteria determined by different measures should be emphasized in attribute reduction.

Generally speaking, if the definition of attribute reduction is given, then the immediate problem is to find the reduct. In the field of rough sets, widely accepted strategies for finding reducts include the exhaustive and heuristic algorithms. Although the former can be used to derive all reducts, it is not suitable for practical applications because of huge computational complexity. Different from the exhaustive algorithm, the heuristic algorithm can be realized by a greedy searching strategy, which has been favored by most researchers thanks to its speed advantage.

Nevertheless, it is worth noting that the process of the heuristic algorithm is still based on scanning all samples in the data. For example, to generate the binary relation that characterizes the similarity between any two samples, all data samples must be taken into account. However, time consumption may fail to satisfy practical applications with the sharply increasing of the size of dataset. In order to tackle this problem, various attribute-reduction algorithms have been proposed that intended to utilize sample selection to reduce the dataset and simultaneously maintain the performance of the classification accuracy. Angiulli et al. [13] proposed the Fast Condensed Nearest Neighbor, which attempt to remove surplus samples and reduce the dataset. Li et al. [14] presented a critical pattern selection algorithm by considering local geometrical and statistical information. This algorithm selected both border and edge patterns from the dataset. Nicolia et al. [15] proposed a sample selection algorithm dealing with the class-imbalance problem. Lin et al. [16] proposed an approach for detecting the representative samples from large datasets. Zhai et al. [17,18] have analyzed the algorithms above and proposed the sample selection approaches that aim to achieve the same performance of a machine-learning algorithm as the whole dataset is used. Zhang et al. [19] proposed a fuzzy rough set-based information entropy for sample selection in a mixed dataset. Xu et al. [20] have applied the sample selection technique to multilabel feature selection for reducing time consumption. Following these results, sample selection may be feasible and effective.

From the above discussions, the motivation of this paper was to design a new framework of attribute reduction that considers both multiple criteria in the definition of attribute reduction and sample selection in the process of finding reducts. Firstly, sample selection was executed, which aims to decrease the number of samples. Immediately, lower time consumption may be achieved, mainly because the size of the data has been reduced. Secondly, the multicriteria strategy that considers at least two measures was designed. Finally, the voting mechanism was used to select the candidate attribute in each iteration, and then the multiple criteria reduct with sample selection was obtained. Our approach can not only guarantee that the derived reduct satisfies the constraints in terms of different measures, but also make the ensemble selection [5,21] of attributes possible.

The rest of this paper is organized as follows. First, we review some basic concepts in Section 2.1. In Section 3, following the limitations of a single measure, a multiple criteria-based attribute reduction strategy is studied. Then, sample selection and the multiple-criteria strategy are combined to find the corresponding reduct. In Section 4, the effectiveness of our approach over 12 UCI datasets is analyzed. We then conclude with some remarks and perspectives for future work in Section 5.

## 2. Preliminaries

### 2.1. Neighborhood Relation

Without loss of generality, a decision system can be described as  $DS = \langle U, AT \cup \{d\} \rangle$ , in which universe  $U$  is the set of samples,  $AT$  is the set of condition attributes, and  $d$  is a decision attribute. Furthermore,  $\forall x \in U$ ,  $d(x)$  indicates the label or decision value of sample  $x$ , and  $a(x)$  denotes the value of  $x$  over condition attribute  $a$  where  $a \in AT$ .

Given decision system  $DS$ , assume that all decision values in  $DS$  are discrete, and an indiscernibility relation [22–24]  $IND_d$  can be defined as:

$$IND_d = \{(x, y) \in U \times U : d(x) = d(y)\}. \quad (1)$$

By  $IND_d$ , we can obtain a partition over the universe, such that

$$U/IND_d = \{X_1, X_2, \dots, X_q\}, \quad (2)$$

i.e., universe  $U$  is partitioned into  $q$  different decision classes. Therefore,  $\forall X_i \in U/IND_d$ ,  $X_i$  is referred to as the  $i$ -th decision class in rough-set theory. The decision class that contains the sample  $x$  is denoted by  $[x]_d$ .

The rough-set objective is to approximate the decision classes by the information given by condition attributes. Such information can actually be represented by the form of information granules [25,26] from the viewpoint of granular computing. For instance, the equivalence class used in traditional rough sets is a typical example of an information granule.

Nevertheless, it should be emphasized that the equivalence classes are only suitable for dealing with categorical data, while numerical data [27–29] have been seen everywhere in real-world applications. To fill such a gap, many different types of information granules have been proposed. As an important information granule used in generalized rough sets, neighborhood has been widely accepted by researchers. This is mainly because: (1) the construction of neighborhood is based on the distance that can characterize the similarity between samples with numerical data; (2) different neighborhood scales can be easily obtained by using different radii, and then a multigranularity structure is naturally formed. To know what neighborhood is, the concept of neighborhood relation should be given as follows:

Given a decision system  $DS$ ,  $\forall A \subseteq AT$ , suppose that  $\Delta_A : U \times U \in \mathbb{R}^+ \cup \{0\}$  is the Euclidean distance function in which  $\mathbb{R}^+$  is the set of positive real numbers, then  $\Delta_A(x, y)$  represents the Euclidean distance between samples  $x$  and  $y$  by using the information over condition attributes in  $A$ . Immediately, the neighborhood relation is:

$$N_A = \{(x, y) \in U \times U : \Delta_A(x, y) \leq \delta\}, \quad (3)$$

in which  $\delta$  is a given radius such that  $\delta \geq 0$ .

Based on neighborhood relation, it is not difficult to obtain the neighborhood of  $x$  in terms of  $A$ , such that:

$$\delta_A(x) = \{y \in U : \Delta_A(x, y) \leq \delta\}. \quad (4)$$

## 2.2. Neighborhood Rough Set and Neighborhood Classifier

Given a sample  $x$ , to avoid that only the sample  $x$  is in the neighborhood of  $x$ , which may bring us the difficulty for classification, Hu et al. [30] have modified the radius such that

$$\delta' = \min_{y \in U - \{x\}} \Delta_A(x, y) + \delta \cdot \left( \max_{y \in U - \{x\}} \Delta_A(x, y) - \min_{y \in U - \{x\}} \Delta_A(x, y) \right). \quad (5)$$

Following Equation (5), the modified neighborhood of sample  $x$  in terms of  $A$  is

$$\delta'_A(x) = \{y \in U : \Delta_A(x, y) \leq \delta'\}. \quad (6)$$

**Definition 1.** Given decision system  $DS$ ,  $\forall A \subseteq AT$  and  $\forall X_i \in U/IND_d$ , the neighborhood lower and upper approximations of  $X_i$  in terms of  $A$  are defined as

$$\underline{X}_{iA} = \{x \in U : \delta'_A(x) \subseteq X_i\}; \quad (7)$$

$$\overline{X}_{iA} = \{x \in U : \delta'_A(x) \cap X_i \neq \emptyset\}. \quad (8)$$

Pair  $[\underline{X}_{iA}, \overline{X}_{iA}]$  is referred to as a neighborhood rough set of  $X_i$  in terms of  $A$ .

The concept of neighborhood can not only be used to construct a rough set, but can also be applied to design classifier [31]. Let us consider one of the simplest classifiers, i.e., the  $K$  Nearest Neighbors algorithm (KNN), which is effective in many cases. It is a lazy learning method: for a testing sample to be classified, its  $K$  nearest neighbors form a neighborhood of such testing sample, and the voting mechanism is used to determine the label of the testing sample based on the real labels of all samples in neighborhood. For more details about KNN algorithm, see references [32–34].

The main thinking of the neighborhood classifier [8] is similar to that of KNN; the difference lies in the fact that the number of neighbors used in a neighborhood classifier is determined by the radius, while the number of neighbors used in KNN is specified by experts. Therefore, different samples may have different numbers of neighbors if a neighborhood classifier is used. The detailed process of the neighborhood classifier [8] is shown in Algorithm 1, as follows.

---

**Algorithm 1** Neighborhood Classifier (NEC)

---

**Inputs:** Decision system  $DS = \langle U, AT \cup \{d\} \rangle$ , a testing sample  $y \notin U$ , radius  $\delta$ ;

**Outputs:** Predicted label of  $y$ :  $\text{Pre}_{AT}(y)$ .

1.  $\forall x \in U$ , compute  $\Delta_{AT}(y, x)$ ;
  2. Compute  $\delta'$  by Equation (5), and then obtain  $\delta'_{AT}(y)$  by Equation (6);  
// Note that in NEC,  $y \notin \delta'_{AT}(y)$ ;
  3.  $\forall X_i \in U/IND_d$ , compute the probability  $\Pr(X_i | \delta'_{AT}(y)) = \frac{|\delta'_{AT}(y) \cap X_i|}{|\delta'_{AT}(y)|}$ ;
  4.  $X_k = \arg \max \{\Pr(X_i | \delta'_{AT}(y)) : \forall X_i \in U/IND_d\}$ ;
  5. Find the corresponding label in terms of  $X_k$  and assign it to  $\text{Pre}_{AT}(y)$ ;
  6. Return  $\text{Pre}_{AT}(y)$ .
- 

### 2.3. Measures

Approximation quality is frequently used to evaluate the certainty of belongingness in rough-set theory. In a neighborhood rough set, the formal definition is shown as follows:

**Definition 2.** Given decision system  $DS$ ,  $\forall A \subseteq AT$ , the approximation quality of  $d$  related to  $A$  is defined as

$$\gamma(A, d) = \frac{|\bigcup_{i=1}^q \underline{X}_{iA}|}{|U|}. \quad (9)$$

This reflects the percentage of the samples that belong to one of the decision classes determinately by the semantic explanation of lower approximation. Obviously,  $0 \leq \gamma(A, d) \leq 1$  holds.

**Remark 1.** Note that  $\forall A \subseteq AT$ ,  $\gamma(A, d) \leq \gamma(AT, d)$  does not always hold; the reason is that, if some condition attributes are eliminated from  $AT$ , then the value of  $\delta'$  obtained by Equation (5) changes.

**Example 1.** As the decision system shown in Table 1,  $U = \{x_1, x_2, \dots, x_{10}\}$  is the set of samples,  $AT = \{a_1, a_2, \dots, a_7\}$  is the set of condition attributes, and  $d$  is a decision attribute.

Suppose that  $\delta = 0.10$ , then by Equation (5), the values of  $\delta'$  for ten samples are 0.8553, 0.7208, 0.8463, 0.8583, 0.8651, 0.8183, 0.7941, 0.6669, 0.9005, and 0.7025, respectively, if  $AT$  is used. Consequently, it is obtained that  $\gamma(AT, d) = 0.1000$ .

Suppose that condition attribute  $a_7$  is eliminated from the above decision system and then  $A = \{a_1, a_2, \dots, a_6\}$ , by Equation (5), the values of  $\delta'$  for ten samples are 0.8212, 0.7059, 0.7633, 0.8349, 0.6992, 0.6860, 0.6738, 0.6289, 0.8439, and 0.6767, respectively. Immediately,  $\gamma(A, d) = 0.2000$  is obtained.

The above results tell us that  $\gamma(A, d) \leq \gamma(AT, d)$  does not always hold if  $A \subseteq AT$ .

**Table 1.** Decision-system example.

	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$a_6$	$a_7$	$d$
$x_1$	0.8147	0.1576	0.6557	0.7060	0.4387	0.2760	0.7513	1
$x_2$	0.9058	0.9706	0.0357	0.0318	0.3816	0.6797	0.2551	1
$x_3$	0.1270	0.9572	0.8491	0.2769	0.7655	0.6551	0.5060	2
$x_4$	0.9134	0.4854	0.9340	0.0462	0.7952	0.1626	0.6991	2
$x_5$	0.6324	0.8003	0.6787	0.0971	0.1869	0.1190	0.8909	3
$x_6$	0.0975	0.1419	0.7577	0.8235	0.4898	0.4984	0.9593	3
$x_7$	0.2785	0.4218	0.7431	0.6948	0.4456	0.9597	0.5472	1
$x_8$	0.5469	0.9157	0.3922	0.3171	0.6463	0.3404	0.1386	2
$x_9$	0.9575	0.7922	0.6555	0.9502	0.7094	0.5853	0.1493	3
$x_{10}$	0.9649	0.9595	0.1712	0.0344	0.7547	0.2238	0.2575	2

Conditional entropy is another widely accepted measure that is an effective tool for characterizing distinguishable ability in a decision system. The lower the value of conditional entropy is, the higher the ability of that condition attribute to distinguish samples that we will have. Such discrimination can be considered as a type of uncertainty. Presently, many definitions of conditional entropies have been proposed in terms of different requirements [9–11,35–37]. A typical representation of conditional entropy is shown in Definition 3.

**Definition 3.** Reference[30] Given decision system  $DS$ ,  $\forall A \subseteq AT$ , the conditional entropy of  $d$  related to  $A$  is defined as:

$$\text{ENT}(A, d) = -\frac{1}{|U|} \sum_{x \in U} \log \frac{|\delta'_A(x) \cap [x]_d|}{|\delta'_A(x)|}. \quad (10)$$

$\forall A \subseteq AT$ ,  $\text{ENT}(A, d) \geq \text{ENT}(AT, d)$  does not always hold. This is because not only does the monotonic property of Equation (10) not hold [19], but also the value of  $\delta'$  obtained from Equation (5) is changed if some condition attributes are eliminated from  $AT$ .

### 3. Multiple-Criteria Reduct with Sample Selection

#### 3.1. Attribute Reduction

Attribute reduction is one of the key topics in rough-set theory [38]. Generally speaking, the purpose of attribute reduction is to delete the redundant attributes by some given constraints [39]. These constraints can be constructed by well-known measures such as approximation quality and conditional entropy. Many different definitions of attribute reduction have been proposed with different measures or requirements. Dai et al [9] proposed extended conditional entropy in interval-valued decision systems and designed corresponding definitions and algorithms. Yao et al. [7] addressed different measures, such as confidence, coverage, generality, cost, and decision monotocity based on the decision-theoretic rough-set models. Jia et al. [40] compared most popular definitions and then proposed a generalized attribute reduction that not only considers the data but also users' preferences. For more details about definitions of attribute reductions, see references [9,41–43].

**Definition 4.** Given decision system  $DS$ ,  $\forall A \subseteq AT$ ,

1.  $A$  is the approximation quality reduct if and only if  $\gamma(A, d) \geq \gamma(AT, d)$  and  $\forall A' \subset A$ ,  $\gamma(A', d) < \gamma(A, d)$ ;
2.  $A$  is the conditional entropy reduct if and only if  $ENT(A, d) \leq ENT(AT, d)$  and  $\forall A' \subset A$ ,  $ENT(A', d) > ENT(A, d)$ .

Different from Pawlak's [1,2] traditional definition of attribute reduction for preserving approximation quality, the constraint used in Definition 4 indicates that the approximation quality will not be decreased at least. The reason is shown in Remark 1: the approximation quality is not strictly monotonic in terms of variations of condition attributes. The case of conditional entropy is similar to that of approximation quality.

If  $\gamma(A \cup \{a\}, d) \leq \gamma(A, d)$ , then  $a$  is redundant; in other words, attribute  $a$  has no contribution to the increase of approximation quality. If  $\gamma(A \cup \{a\}, d) > \gamma(A, d)$ , then  $a$  can be considered as a member in the reduct set. Similarly, it is trivial to present the semantic explanation of redundant attributes in terms of conditional entropy reduct. Therefore, the significance of attributes in terms of two different reducts shown in Definition 4 can be defined as follows:

$$\text{Sig}_\gamma(a, A, d) = \gamma(A \cup \{a\}, d) - \gamma(A, d); \quad (11)$$

$$\text{Sig}_{\text{ENT}}(a, A, d) = ENT(A, d) - ENT(A \cup \{a\}, d). \quad (12)$$

In a decision system, the above two significances both satisfy that the higher the value is, the more important the condition attribute  $a$  will be. Following the given significances, the algorithms of finding the reduct must be immediately designed. Up to now, many algorithms have been proposed to obtain reducts. Considering time efficiency, the forward greedy search strategy has become a common way to do this. This kind of algorithm starts from an empty set and gradually adds the attribute with the maximum significance into the candidate attribute subset in each iteration [44] until the constraint is satisfied. This kind of approach is frequently referred to as the heuristic algorithm.

Take the approximation quality reduct as an example; the reduct aims to derive a subset of condition attributes that do not decrease the value of approximation quality. The detailed process of heuristic algorithm for finding such reduct is shown in Algorithm 2.

---

**Algorithm 2** Approximation Quality Reduct (AQR)

---

**Inputs:** Decision system  $DS = \langle U, AT \cup \{d\} \rangle$ , radius  $\delta$ .

**Outputs:** An approximation quality reduct  $A$ .

---

1. Compute  $\gamma(AT, d)$ ;
  2.  $A \leftarrow \emptyset$ ;
  3. **Do**
    - (1)  $\forall a_i \in AT - A$ , compute  $\text{Sig}_\gamma(a_i, A, d)$ , select  $a_j$  such that  $\text{Sig}_\gamma(a_j, A, d) = \max\{\text{Sig}_\gamma(a_i, A, d) : \forall a_i \in AT - A\}$ ;
    - (2)  $A = A \cup \{a_j\}$ ;
    - (3) Compute  $\gamma(A, d)$ ;**Until**  $\gamma(A, d) \geq \gamma(AT, d)$ ;
  4. **Return**  $A$ .
- 

Similarly, if it is required to compute Conditional Entropy Reduct (CER), as the conditional entropy is a measure that characterizes the distinguishing information of a subset of attributes, and the lower of the value of the conditional entropy is, the greater the distinguishment ability of the attribute set is. Then, the termination in Step 3 of Algorithm 2 is replaced by " $ENT(A, d) \leq ENT(AT, d)$ ", and the significance of attribute in Step 3(1) is replaced by Equation (12), i.e.,  $\text{Sig}_{\text{ENT}}(a_i, A, d)$ ; then, we select the attribute that has maximum significance.



The time complexity of a computing neighborhood relation is  $O(|U|^2)$ , in which  $|U|$  is the number of samples in a dataset. In the worst case, there are  $|AT|$  attributes should be added into the reduct, i.e., no attribute is redundant; then, Step 3 in Algorithm 2 is executed  $|AT|$  times. In the  $i$ -th iteration, Step 3 is executed  $|AT| - i + 1$  times. Finally, the time complexity of AQR is  $O(|U|^2 \times |AT|^2)$ . Similarly, the time complexity of CER is also  $O(|U|^2 \times |AT|^2)$ .

### 3.2. Limitations of Single Measure

The above algorithm shows us a complete process of computing the reduct that is determined by a single measure, i.e., either approximation quality or conditional entropy. However, the derived reduct may fail to meet the constraints with multiple criteria. We used the following example to explain it:

**Example 2.** In the decision system shown in Table 1, suppose that  $\delta = 0.15$ , then  $\gamma(AT, d) = 0.1000$  and  $ENT(AT, d) = 0.6879$ ; both of them are obtained by raw attributes.

Furthermore, by Definition 4 and the heuristic process, the obtained approximation-quality reduct is  $A_1 = \{a_1, a_2, a_3\}$ , and the obtained conditional-entropy reduct is  $A_2 = \{a_1, a_2, a_3, a_4, a_5\}$ .

If the approximation-quality reduct is selected, then  $\gamma(A_1, d) = 0.1000$ , and  $ENT(A_1, d) = 0.7860$ . It is observed that the value of approximation quality is maintained, while the value of conditional entropy is increased. In other words, though the constraint based on approximation quality is satisfied by  $A_1$ , such subset of attributes does not meet the constraint in terms of conditional entropy.

If the conditional-entropy reduct is selected, then  $\gamma(A_2, d) = 0.0000$ , and  $ENT(A_2, d) = 0.6762$ . It is observed that the value of conditional entropy has been significantly decreased, while the value of approximation quality is also decreased. In other words, the constraint defined by approximation quality cannot be guaranteed if the conditional-entropy reduct is used.

The above example tells us that a reduct in terms of a single measure does not meet the constraints in terms of multiple criteria. Therefore, to alleviate such a problem, we propose a multiple-criteria attribute-reduction strategy that considers both the evaluations of approximation quality and conditional entropy.

### 3.3. Multiple-Criteria Reduct

Since the single-criterion reduct cannot meet the constraints, then the multiple-criteria framework [45] can be a solution. The definition of a multiple-criteria reduct presented as follows:

**Definition 5.** Given decision system  $DS$ , if  $A \subseteq AT$ ,  $A$  is the multiple-criteria reduct if and only if:

1.  $\gamma(A, d) \geq \gamma(AT, d)$  and  $ENT(A, d) \leq ENT(AT, d)$ ;
2.  $\forall A' \subset A$ ,  $\gamma(A', d) < \gamma(A, d)$  or  $ENT(A', d) > ENT(A, d)$ .

Different from the approximation-quality and conditional-entropy reducts shown in Definition 4, the multiple-criteria reduct shown in Definition 5 is defined by considering constraints given by both approximation quality and conditional entropy.

Algorithm 3 presents a heuristic process to compute our multiple-criteria reduct. It should be emphasized that, to derive attribute significance, Equations (11) and (12) should be used.

In Step 3,  $m$  and  $n$  represent the attribute locations that have maximal significances in terms of approximation quality and conditional entropy, respectively. In Step 3(2), if  $m$  and  $n$  are the same, then there is no conflict for voting. Otherwise, two attributes have conflict, which means that the max values of significances computed by the measures of approximation quality and conditional entropy are derived from different attributes. Then, a mechanism for selection is required. In this case, we select one attribute without considering the measures, mainly because approximation quality and conditional entropy take the same weight in our algorithm. To make the algorithm more stable,

the attribute ranks lower in the order of the raw attributes that are selected instead of a random one. Such thinking is similar to what has been addressed in reference [21].

Similar to AQR, the time complexity of MCR is  $O(|U|^2 \times |AT|^2)$ , where  $|U|$  represents the number of samples in decision system ( $DS$ ), and  $|AT|$  represents the number of condition attributes. However, MCR may spend more time on computing reduct, because MCR should compute two different types of significances in each iteration.

---

**Algorithm 3** Multiple-Criteria Reduct (MCR)
 

---

**Inputs:** Decision system  $DS = \langle U, AT \cup \{d\} \rangle$ , radius  $\delta$ .

**Outputs:** A multiple criteria reduct  $A$ .

1. Compute  $\gamma(AT, d)$  and  $ENT(AT, d)$ ;
  2.  $A \leftarrow \emptyset$ ;
  3. **Do**
    - (1)  $\forall a_i \in AT - A$ , compute  $Sig_\gamma(a_i, A, d)$  and  $Sig_{ENT}(a_i, A, d)$ , select  $a_m$  and  $a_n$  such that  
 $Sig_\gamma(a_m, A, d) = \max\{Sig_\gamma(a_i, A, d) : \forall a_i \in AT - A\}$ ,  
 $Sig_{ENT}(a_n, A, d) = \max\{Sig_{ENT}(a_i, A, d) : \forall a_i \in AT - A\}$ ;
    - (2) Select  $a_j$ , where  $j = \min(m, n)$ ;
    - (3)  $A = A \cup \{a_j\}$ ;
    - (4) Compute  $\gamma(A, d)$  and  $ENT(A, d)$ ;
  - Until**  $\gamma(A, d) \geq \gamma(AT, d)$  and  $ENT(A, d) \leq ENT(AT, d)$ ;
  4. Return  $A$ .
- 

### 3.4. Multiple-Criteria Reduct with Sample Selection

Obviously, the process of the heuristic algorithm for computing the reduct is still based on scanning all samples in the data. To further improve the time efficiency of the algorithms shown in Sections 3.1 and 3.3, reducing the size of samples may be a feasible solution.

In the field of machine learning and feature selection [46,47], the technique of sample selection has been widely used. For instance, following many previous results [20,48–50], it has been pointed out that sample selection is a useful method. Wilson et al. [48] provided a survey of previous algorithms, and proposed six additional reduction algorithms that can be used to remove samples from the concept description. Brighton et al. [49] proposed that internal samples positioned away from class boundaries have little or no effect on classification accuracy; on the contrary, samples that lie close to class boundaries hold more information to accurately describe the decision surface. Nikolaidis et al. [50] proposed the Class Boundary Preserving Algorithm (CBP). CBP divided all data into two sets that are referred to as the internal-sample set and boundary-sample set, and focused more on the boundary samples. Xu et al. [20] further expanded the sample selection of boundary samples and introduced it into multilabel datasets. From the above analyses, we can find that the samples in the boundary region are more important than other samples. We propose an algorithm to compute a multiple-criteria reduct by using boundary samples instead of whole samples in the data.

First of all, we used a  $K$ -means clustering algorithm to choose  $K$  cluster centroids [51–54]. This process is executed  $M$  times because the result of  $K$ -means clustering is not stable. Secondly, we compute average cluster centroids by those results. Finally, we select those samples that are far away from the average cluster centroids, and construct a new decision system. We select those samples that are far away from the average cluster centroids, which is mainly because: (1) these samples are more difficult to be correctly classified, and samples nearer to the average cluster centroids tend to be closer to each other, making it easy for them to be classified correctly; (2) these samples sometimes fail to be assigned to the lower approximation set, while the samples closer to the average cluster centroids tend to be in the lower approximation set. Therefore, in order to improve classification performance and reduce the time consumption with the neighborhood rough-set model, we apply boundary samples instead of all samples. To judge whether a sample is far away from the cluster centroid, we used Definition 6, as follows:



**Definition 6.** Given a cluster  $C_j$ ,  $C_j^*$  is the cluster centroid of  $C_j$ ,  $\text{dist}(x, C_j^*)$  denotes the distance between  $x \in C_j$  and  $C_j^*$ , and the average distance between all samples in  $C_j$  and  $C_j^*$  is

$$\overline{\text{dist}}(C_j^*) = \frac{1}{|C_j|} \sum_{x \in C_j} \text{dist}(x, C_j^*). \quad (13)$$

**Remark 2.**  $\forall x \in C_j$ , if  $\text{dist}(x, C_j^*) \geq \overline{\text{dist}}(C_j^*)$ , then  $x$  is referred to as a sample which is far away from the cluster centroid  $C_j^*$ , such sample is selected for constructing new decision system.

With all boundary samples selected, new decision system  $DS'$  can be constructed. Obviously, the size of the data in  $DS'$  is smaller than that in decision system  $DS$ . From this point of view, the time consumption of computing the reduct may be reduced. Algorithm 4 shows us the heuristic process to compute a multiple-criteria reduct by using sample selection.

---

**Algorithm 4** Multiple-Criteria Reduct with Sample Selection (MCRSS)

---

**Inputs:** Decision system  $DS = \langle U, AT \cup \{d\} \rangle$ , radius  $\delta$ ,  $M$ .

**Outputs:** A multiple criteria reduct  $A$ .

1.  $U' = \emptyset$ ;  
// Initialize the universe of new decision system;
2. **For**  $r = 1$  to  $M$   
    Execute  $K$ -means clustering algorithm over  $DS$ , obtain clusters  $C^r = \{C_1^r, C_2^r, \dots, C_K^r\}$ ;  
    // In  $K$ -means clustering,  $K$  is the number of decision classes;  
    **End For**
3. **For**  $j = 1$  to  $K$   
    Obtain the  $j$ -th average cluster centroid

$$C_j^* = \frac{\sum_{r=1}^M C_j^r}{M};$$

- End for**
  4. **For**  $j = 1$  to  $K$   
     $\forall x \in C_j$ , if  $\text{dist}(x, C_j^*) \geq \overline{\text{dist}}(C_j^*)$ , then  $U' = U' \cup \{x\}$ ;  
    **End For**  
    // The new decision system  $DS' = \langle U', AT \cup \{d\} \rangle$  is constructed;
  5. Compute  $\gamma(AT, d)$  and  $\text{ENT}(AT, d)$  over  $DS'$ ;
  6.  $A \leftarrow \emptyset$ ;
  7. **Do**
    - (1)  $\forall a_i \in AT - A$ , compute  $\text{Sig}_{\gamma}(a_i, A, d)$  and  $\text{Sig}_{\text{ENT}}(a_i, A, d)$ , select  $a_m$  and  $a_n$  such that  
 $\text{Sig}_{\gamma}(a_m, A, d) = \max\{\text{Sig}_{\gamma}(a_i, A, d) : \forall a_i \in AT - A\}$ ,  
 $\text{Sig}_{\text{ENT}}(a_n, A, d) = \max\{\text{Sig}_{\text{ENT}}(a_i, A, d) : \forall a_i \in AT - A\}$ ;
    - (2) Select  $a_j$ , where  $j = \min(m, n)$ ;
    - (3)  $A = A \cup \{a_j\}$ ;
    - (4) Compute  $\gamma(A, d)$  and  $\text{ENT}(A, d)$ ;**Until**  $\gamma(A, d) \geq \gamma(AT, d)$  and  $\text{ENT}(A, d) \leq \text{ENT}(AT, d)$ ;
  8. **Return**  $A$ .
- 

The first four steps show us the process of sample selection, i.e., the process of constructing a new decision system. In Step 1, the universe of the new decision system is initialized. In the following two steps, a  $K$ -means clustering algorithm is executed  $M$  times, and the average cluster centroids are obtained. In Step 4, samples that are far away from the average cluster centroids are immediately selected, and the new decision system is constructed. The last three steps are used to compute a multiple-criteria reduct over the new decision system.

The time complexity of MCRSS is  $O(|U'|^2 \times |AT|^2 + K \times |U| \times M)$ , where  $|U'|$  represents the number of samples in the new decision system ( $DS'$ ); both  $K$  and  $M$  are constants. It must be noted that  $|U'| < |U|$ .

The time complexity of MCR is  $O(|U|^2 \times |AT|^2)$ , so we compare the time complexity between MCR and MCRSS. Generally speaking,  $K \times M \leq (|U| - |U'|^2/|U|) \times |AT|^2$  holds for most of the data because  $K$  and  $M$  are constants that are much less than the number of samples.

Therefore, it is a trivial to show that  $O(|U|^2 \times |AT|^2) \geq O(|U'|^2 \times |AT|^2 + K \times |U| \times M)$  holds, in other words, the time consumption of MCRSS is less than that of MCR.

The sample-selection strategy shown in Algorithm 4 can also be used in computing approximation-quality and conditional-entropy reducts. It is immediately trivial to design two algorithms: Approximation-Quality Reduct with Sample Selection (AQRSS) and Conditional-Entropy Reduct with Sample Selection (CERSS). The time complexities of AQRSS and CERSS are also  $O(|U'|^2 \times |AT|^2 + K \times |U| \times M)$ .

#### 4. Experimental Analysis

To validate the effectiveness of MCRSS proposed in this paper, 12 UCI datasets were collected to conduct the experiments. The basic descriptions of the datasets are shown in Table 2. All experiments were carried out on a personal computer with Windows 7, dual-core 1.50 GHz CPU, and 8 GB memory. The programming language was MATLAB R2016a.

**Table 2.** Descriptions of datasets.

ID	Data Sets	Samples	Attributes	Decision Classes
1	Breast Cancer Wisconsin (Diagnostic)	569	30	2
2	Breast Tissue	106	9	6
3	Cardiotocography	2126	21	10
4	Dermatology	365	34	6
5	Forest-Type Mapping	523	27	4
6	Hayes Roth	132	4	3
7	Ionosphere	351	34	2
8	Molecular Biology	106	57	2
9	Statlog (Vehicle Silhouettes)	846	18	4
10	Vertebral Column	310	6	2
11	Wine	178	13	3
12	Yeast	1484	8	10

In the following, five-folder Cross-Validation (5-CV) was adopted. In other words, we divided each set of data into five parts of the same size, which are denoted by  $U_1 \cup \dots \cup U_5$ ; for each round of computation, 80% of the samples in the data were regarded as the training samples for computing reducts, and the rest were considered as the test samples for computing measures by the attributes in reducts. Furthermore, in this experiment,  $M = 5$ , i.e., the  $K$ -means clustering, was executed five times to generate average cluster centroids. Ten different values of  $\delta$ , such that 0.03, 0.06,  $\dots$ , 0.30, were also selected.

##### 4.1. Comparisons of Approximation Qualities

Figure 1 shows us the detailed results of approximation qualities with respect to three different algorithms for computing reducts, AQRSS, MCRSS, and CERSS. AQRSS and CERSS are an approximation-quality reduct and a conditional-entropy reduct with sample selection technique, while MCRSS is a multiple-criteria reduct with sample selection. AQRSS, MCRSS, and CERSS have the same meaning in other comparisons in Section 4.

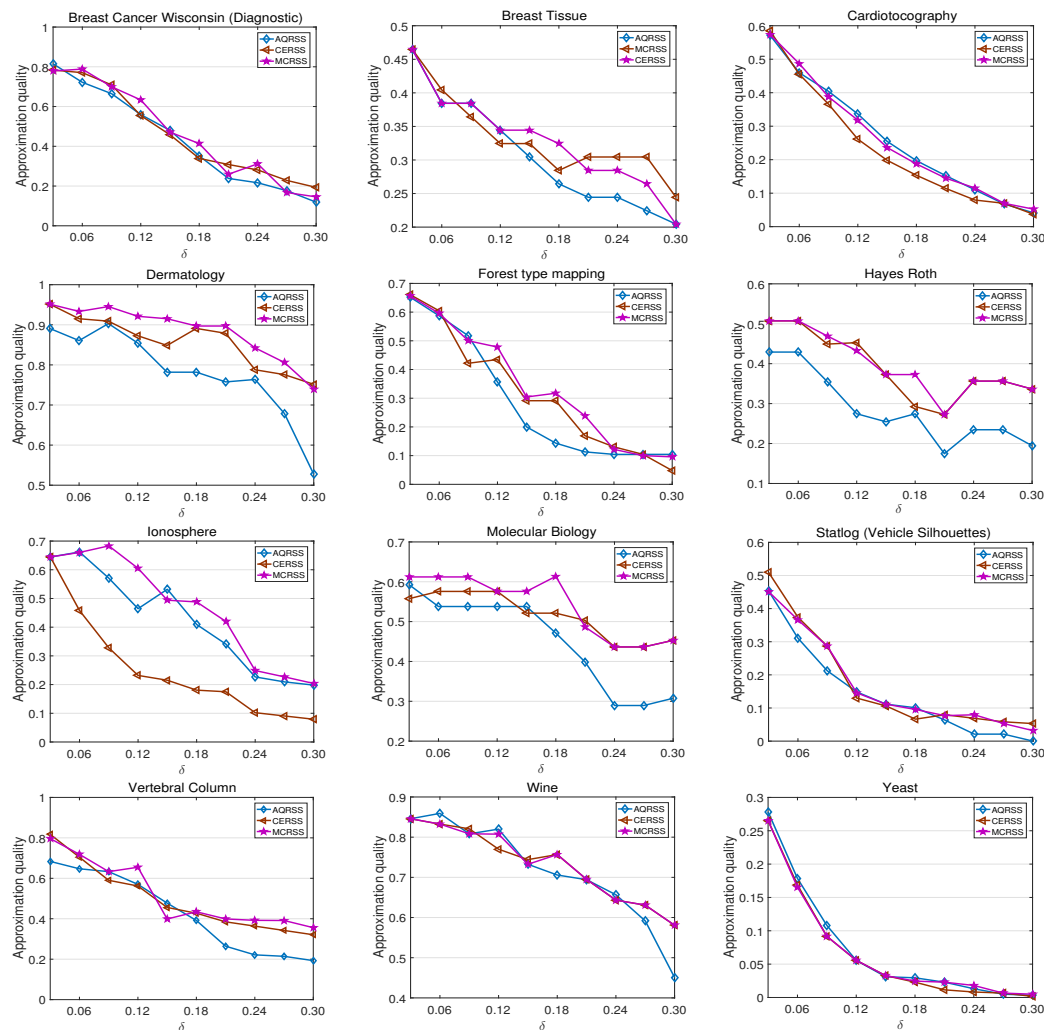


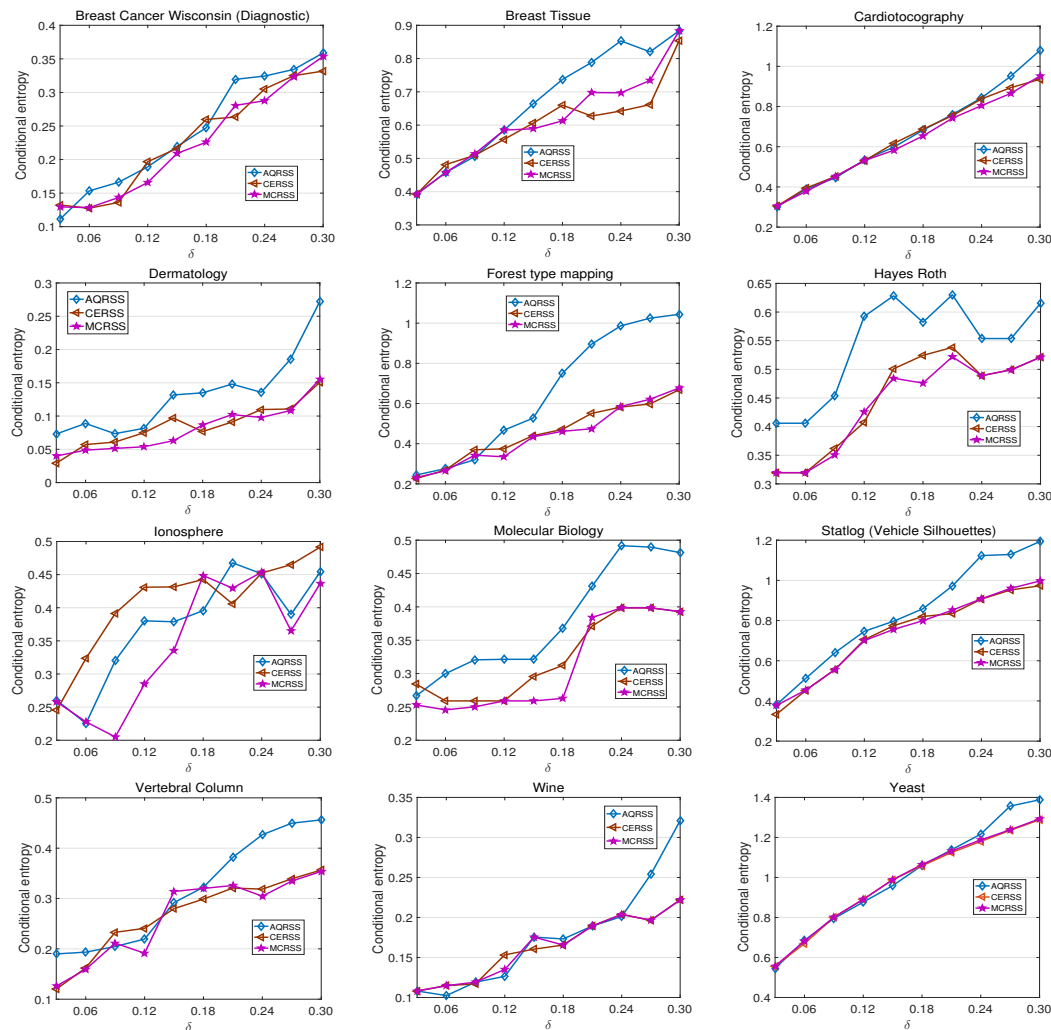
Figure 1. Comparisons of approximation qualities.

In Figure 1, we can observe the following:

1. If the value of  $\delta$  increases, then the decreasing trends have been obtained for approximation qualities with respect to three different reducts, though those decreasing trends are not necessarily monotonic.
2. By comparing it with AQRSS, MCRSS can preserve or slightly increase approximation qualities. This is mainly because the constraint designed by the measure of approximation quality is also considered in MCRSS. Take, for instance, the “Ionosphere” dataset; if  $\delta = 0.12$ , then the approximation qualities derived by MCRSS and AQRSS are 0.6049 and 0.4644, respectively.
3. An interesting observation is that the approximation qualities obtained by CERSS may be greater than those obtained by AQRSS in some datasets. Take, for instance, the “Dermatology” dataset; if  $\delta = 0.06$ , then approximation qualities derived by MCRSS, AQRSS, and CERSS are 0.9333, 0.8606, and 0.9151, respectively. Such results tell us that AQRSS is not always good in deriving higher approximation qualities.

#### 4.2. Comparisons of Conditional Entropies

Figure 2 shows us the detailed results of conditional entropies with respect to three different algorithms for computing reducts.



**Figure 2.** Comparisons of conditional entropies.

In Figure 2, we can observe the following:

1. If the value of  $\delta$  increases, then the increasing trends have been obtained for conditional entropies with respect to three different reducts, though those increasing trends are not strictly monotonic.
2. In most cases, there are slight differences between conditional entropies generated by MCRSS and CERSS, which can be attributed to the constraint designed by the measure of conditional entropy that has also been considered in MCRSS. Take, for instance, the “Breast Tissue” dataset; if  $\delta = 0.15$ , then the conditional entropies derived by MCRSS and CERSS are 0.6127 and 0.6599, respectively.
3. In most cases, the conditional entropies obtained by AQRSS are greater than those derived by both CERSS and MCRSS. This observation demonstrates that, if we only pay attention to the single measure of approximation quality, the obtained reduct may not be effective in terms of conditional entropy. Take, for instance, the “Forest-Type Mapping” dataset; if  $\delta = 0.21$ , then the conditional entropies derived by MCRSS, CERSS, and AQRSS are 0.4744, 0.5507, and 0.8951, respectively.

#### 4.3. Comparisons of Classification Accuracies

In the following, the neighborhood classifier was used to measure the classification performances of the reducts derived from three different algorithms. The detailed results are shown in Figure 3.

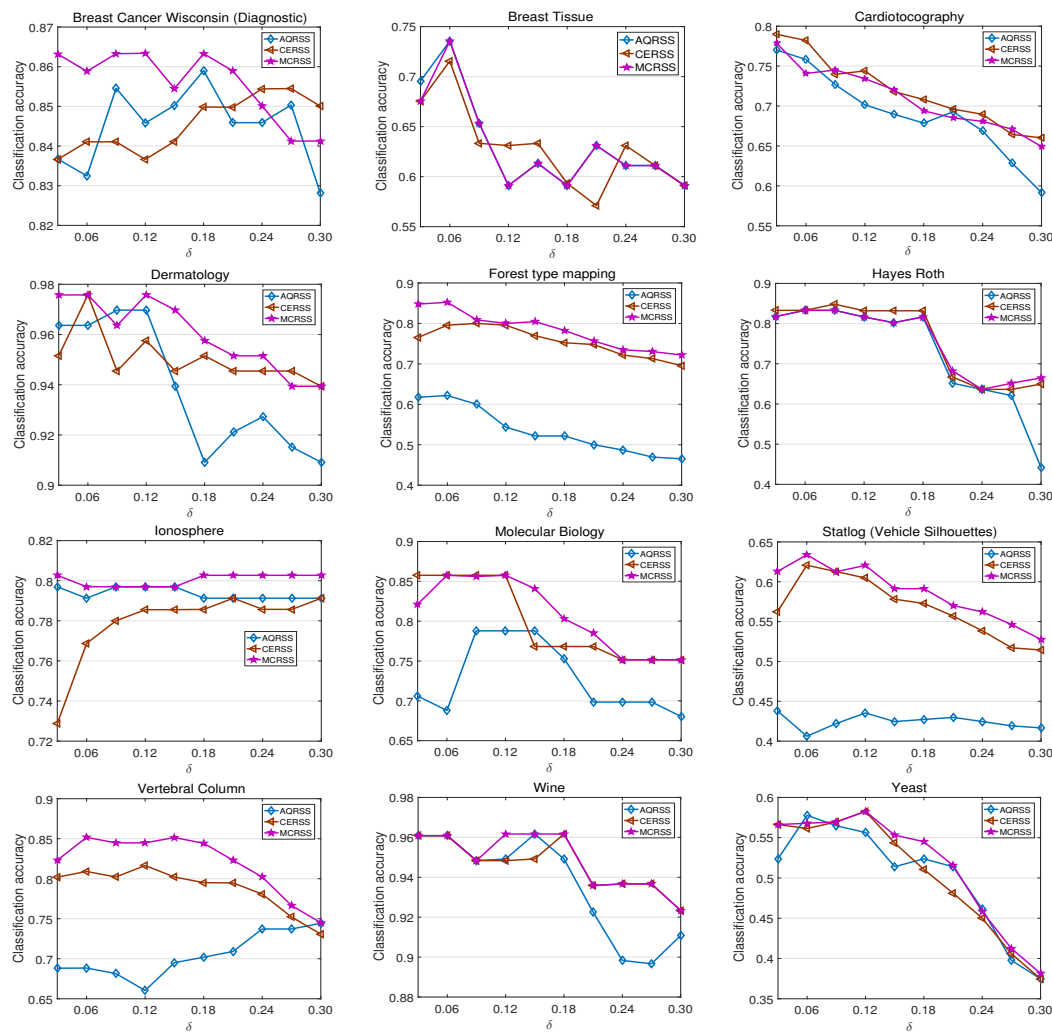


Figure 3. Comparisons of classification accuracies.

In Figure 3, we can observe that the classification accuracies obtained by MCRSS are greater than those obtained by AQRSS and CERSS. Take the “Statlog (Vehicle Silhouettes)” dataset as an example; if  $\delta = 0.06$ , then the classification accuracies derived by MCRSS, AQRSS, and CERSS are 0.6205, 0.4352, and 0.6047, respectively. Such results tell us that the MCRSS algorithm provides better classification performance with the use of a neighborhood classifier.

#### 4.4. Comparisons of Reduct Lengths

Figure 4 shows us the reduct lengths derived from three different algorithms.

In Figure 4, we can observe that the reduct lengths obtained by considering multiple criteria are greater than the lengths of reducts obtained by a single measure (approximation quality or conditional entropy). This is mainly because the multiple-criteria reduct in this experiment considers two measures, and then the constraint is stricter than the constraint defined by only one measure.

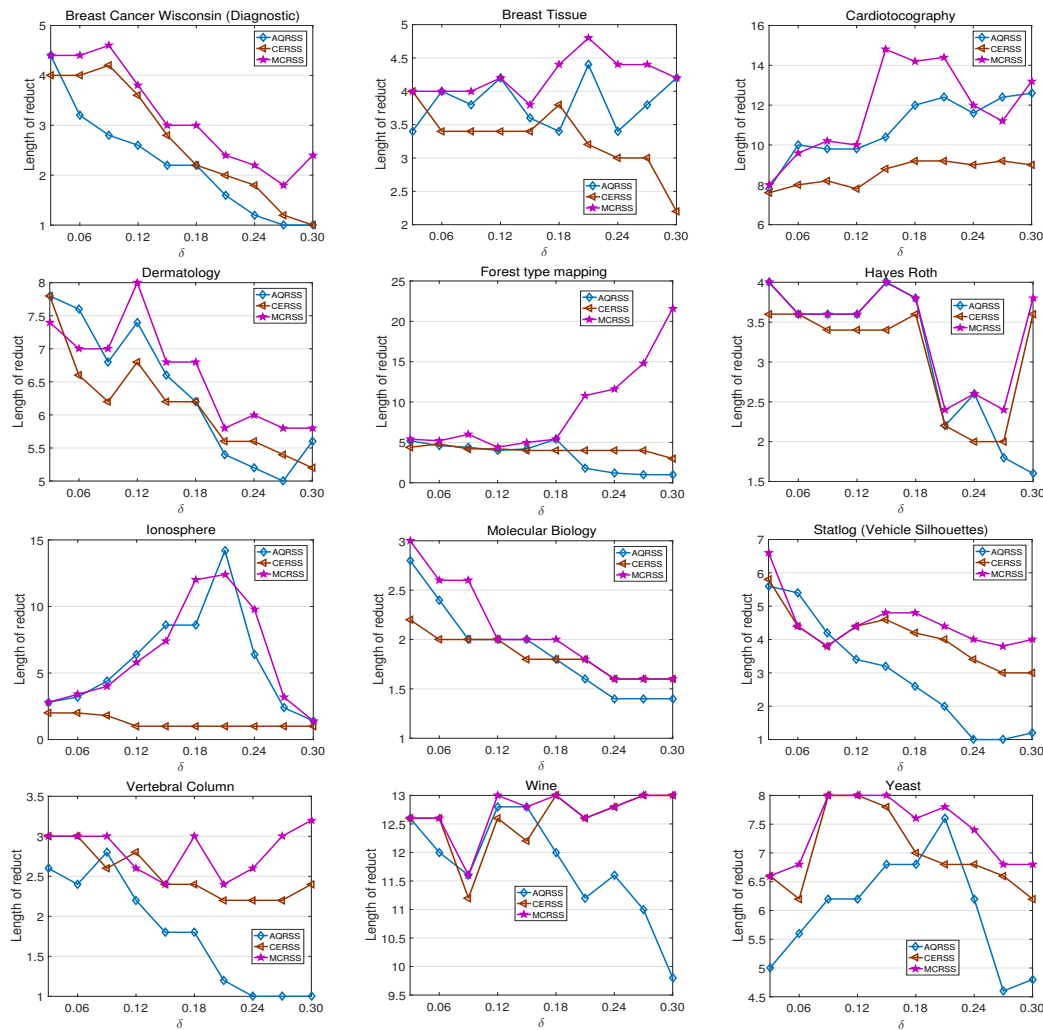


Figure 4. Comparisons of reduct lengths.

#### 4.5. Comparisons of Time Consumptions

In the following, we compare the time consumption of several algorithms, AQRSS, CERSS, MCRSS, and MCR, in generating reducts. AQRSS and CERSS are used to find the approximation-quality reduct and conditional-entropy reduct with sample selection, respectively; MCRSS and MCR are used to find multiple-criteria reducts with and without sample selection, respectively. The detailed results are shown in Figure 5.

The following conclusions can be obtained from Figure 5.

1. The time consumption of MCRSS is higher than that of AQRSS and CERSS, though the time complexities of these three algorithms are the same. The reasons include two aspects: (1) MCRSS computes two attribute significances instead of one in each iteration; (2) the length of the reduct derived by MCRSS is frequently greater than those derived by AQRSS and CERSS, i.e., more iterations should be used.
2. By comparing it with the time consumption of MCR, MCRSS time consumption was significantly reduced. From this point of view, sample selection is effective in the process of finding a reduct from the viewpoint of saving time.



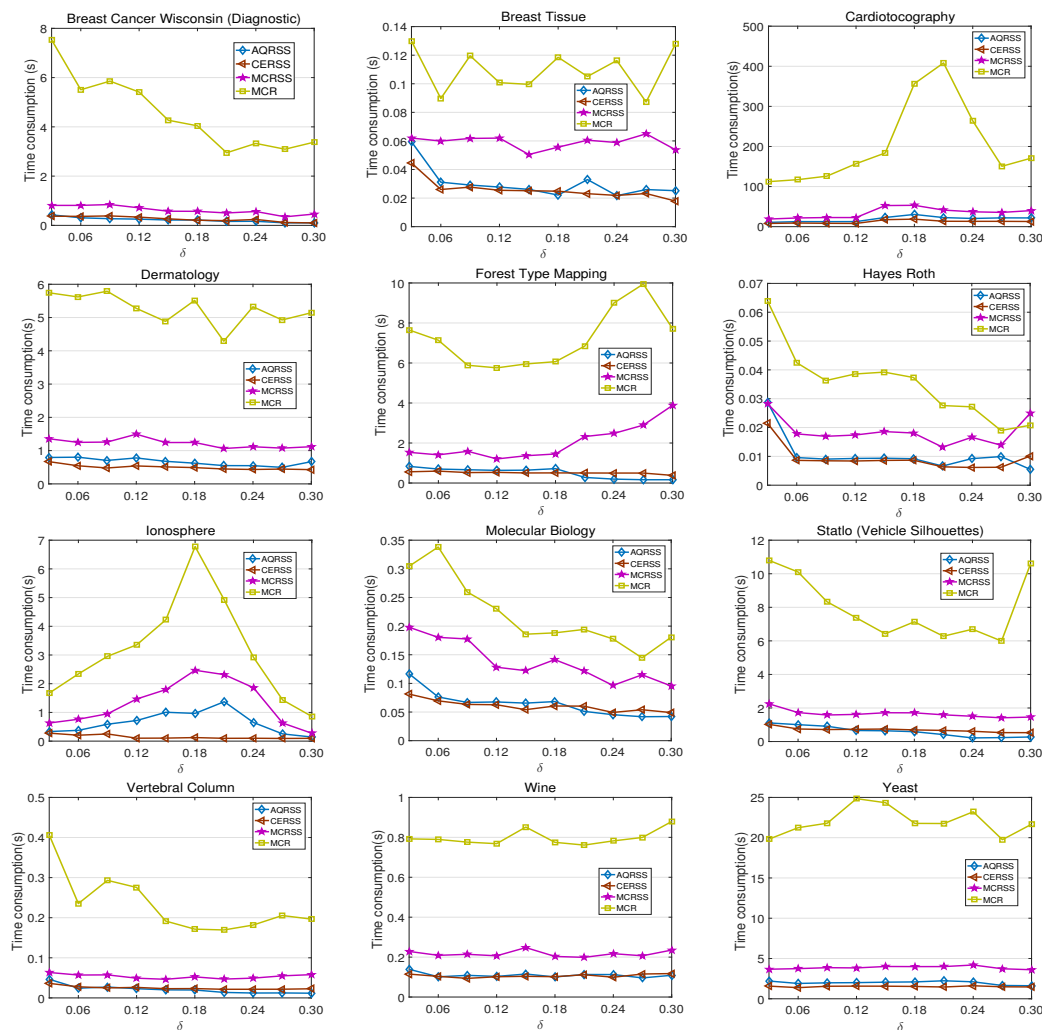


Figure 5. Comparisons of time consumption.

#### 4.6. Comparisons of Core Attributes

In the following, we compare the three algorithms, AQRSS, CERSS and MCRSS, in the view of core attributes. For readers' convenience, we only display the core attributes with one fixed radius; given  $\delta = 0.15$ , we use boundary samples to compute the core attributes [55,56], and the thinking of the process is similar to the algorithm proposed by Wang et al. [56]. We removed only one attribute from the raw attributes ( $AT$ ) to make the subset that is made up of the remaining attributes that cannot satisfy the constraints in definition. Take the measure of "approximation quality" (AQRSS) as an example;  $AT = \{a_1, a_2 \cdots a_n\}$ : (1) remove  $a_1$  in the first time, then the remaining attributes construct the subset  $A = \{a_2, a_3 \cdots a_n\}$ ; (2) compute  $\gamma(AT, d)$  and  $\gamma(A, d)$  in the new decision system  $DS'$ ; (3) if  $\gamma(A, d) < \gamma(AT, d)$ , then  $a_1$  can be a member of the core set.  $a_2$  is removed in the second time and  $a_n$  is removed in the  $n$ -th time. Similar algorithms are used to compute core sets for conditional entropy (CERSS) and multicriterion (MCRSS).

For readers' convenience, to compare the results of these three algorithms (AQRSS, CERSS, MCRSS), the order of attributes was applied. When several consecutive attributes are core attributes, the order are listed with the symbol "-". Take the dataset of "Hayes Roth" (ID: 6) as an example; since core attributes in terms of three algorithms are all  $\{a_1, a_2, a_3, a_4\}$ , which present "1-4" in Table 3. With a careful investigation of Table 3, the core of MCRSS is the union set of the core of AQRSS and CERSS in general. Take the data set of "Forest Type Mapping" (ID: 5) as an example, the core attributes of AQRSS and CERSS are  $\{a_1, a_6, a_7, a_{16}, a_{22}\}$  and  $\{a_1, a_2, a_4, a_6, a_{15}, a_{16}, a_{22}, a_{23}, a_{25}\}$ , respectively. And the core attributes are  $\{a_1, a_2, a_4, a_6, a_7, a_{15}, a_{16}, a_{22}, a_{23}, a_{25}\}$ , which is the union set

of the core attributes of AQRSS and CERSS. In the “Dermatology” (ID: 4) dataset, this union-set relation is not correct since the core attributes of the three algorithms are  $\{a_5, a_{15}\}$ ,  $\{a_5, a_{15}, a_{21}\}$  and  $\{a_5, a_9, a_{15}, a_{21}, a_{22}\}$ , respectively. More information is shown as follows.

**Table 3.** Core attributes.

ID	Datasets	AQRSS	CERSS	MCRSS
1	Breast Cancer Wisconsin (Diagnostic)	2,8,10,28	2,8,20,28,29	2,8,20,28,29
2	Breast Tissue	2-3	2-3,9	2-3,9
3	Cardiotocography	1-2,4-5,8,17,21	2,4-5,8,10-12,18	1-2,4-5,8,10-12,18
4	Dermatology	5,15	5,15,21	5,9,15,21,22
5	Forest-Type Mapping	1,6-7,16,22	1-2,4,6,15-16,22-23,25	1-2,4,6-7,15-16,22-23,25
6	Hayes Roth	1-4	1-4	1-4
7	Ionosphere	1,3,12-14,26,30,32-34	1,3,11,15,17,19,21,32	1,3,11,15,17-22,30,32-34
8	Molecular Biology	13,38-39,48-49,56	13,38-40,48-49	13,38-40,48-49,56
9	Statlog (Vehicle Silhouettes)	4,10,14,16-18	1,4-6,8,10,14,16-18	1,4-6,8,10,14,16-18
10	Vertebral Column	2-3,5	2-3,5	2-3,5
11	Wine	2,5,7,10-13	1-2,4-5,7-8,10-13	1-2,4-5,7-8,10-13
12	Yeast	1-4,6-8	1-7	1-8

Given  $\delta = 0.15$ , the results in Table 4 were obtained from all the data without using sample selection, and 5-CV was also applied to compute the mean values of these three measures (approximation quality, conditional entropy, and classification accuracy).

**Table 4.** Results obtained by using core attributes.

ID	Approximation Quality			Conditional Entropy			Classification Accuracy		
	AQRSS	CERSS	MCRSS	AQRSS	CERSS	MCRSS	AQRSS	CERSS	MCRSS
1	0.6173	<b>0.6303</b>	0.6260	0.1660	0.1661	<u>0.1650</u>	0.9332	0.9560	<b>0.9578</b>
2	0.0800	<b>0.3200</b>	<b>0.3200</b>	0.8963	<u>0.5038</u>	<u>0.5038</u>	0.1229	<b>0.4628</b>	<b>0.4628</b>
3	<b>0.2540</b>	0.2408	0.2352	0.5810	<u>0.5718</u>	0.5784	0.8264	0.8810	<b>0.9045</b>
4	<b>0.9945</b>	0.9918	0.9782	0.1450	0.0896	<u>0.0758</u>	0.5219	0.7760	<b>0.8387</b>
5	0.2783	<b>0.3130</b>	<b>0.3130</b>	0.4716	<u>0.4095</u>	0.4249	0.7342	0.8547	<b>0.8910</b>
6	<b>0.4505</b>	<b>0.4505</b>	<b>0.4505</b>	<u>0.3913</u>	<u>0.3913</u>	<u>0.3913</u>	<b>0.8033</b>	<b>0.8033</b>	<b>0.8033</b>
7	0.6441	0.5646	<b>0.6667</b>	0.4374	<u>0.3442</u>	0.4384	0.7749	<b>0.8319</b>	<b>0.8319</b>
8	<b>0.4030</b>	0.3697	<b>0.4030</b>	<u>0.3919</u>	0.4473	0.4285	0.6710	<b>0.7182</b>	<b>0.7182</b>
9	0.1298	<b>0.1403</b>	<b>0.1403</b>	0.7605	<u>0.7013</u>	<u>0.7013</u>	0.2211	<b>0.5757</b>	<b>0.5757</b>
10	<b>0.4882</b>	<b>0.4882</b>	<b>0.4882</b>	<u>0.2799</u>	<u>0.2799</u>	<u>0.2799</u>	<b>0.8719</b>	<b>0.8719</b>	<b>0.8719</b>
11	0.8142	<b>0.8408</b>	<b>0.8408</b>	0.1085	<u>0.0966</u>	<u>0.0966</u>	0.9665	<b>0.9775</b>	<b>0.9775</b>
12	<b>0.0532</b>	0.0443	0.0443	0.9216	<u>0.9203</u>	<u>0.9203</u>	<b>0.5266</b>	0.5213	0.5213
average	0.4340	0.4495	<b>0.4546</b>	0.4625	<u>0.4101</u>	0.4170	0.6645	0.7713	<b>0.7756</b>

After a careful investigation of Table 4, it can be seen that MCRSS improves approximation quality and classification accuracy, and it also reduces the conditional entropy.

1. In the comparisons of these three measures, the largest values of approximation quality (classification accuracy) are in bold, and the smallest values of conditional entropy are underlined. It should be emphasized that in Datasets 6 and 10, the values of these three measures are the same. This is mainly because the cores of the three algorithms are the same, which can be seen from Table 3.
2. The results shown in Table 4 can generally stay consistent with the results shown above (Figures 1–3), which are obtained from the datasets with sample selection. The reducts obtained by MCRSS can not only preserve approximation quality (Figure 1) and reduce conditional entropy (Figure 2), but also improve classification accuracy performance (Figure 3).
3. We can find that conditional entropy and approximation quality both have an important role in improving performance. The measure of conditional entropy may contribute a little more in improving classification accuracy values. The phenomenon that most of the values derived from

the CERSS and MCRSS are the same may illustrate that the constraint of conditional entropy is more helpful in improving classification accuracy.

## 5. Conclusions and Future Perspectives

In this paper, a framework of a multiple-criteria reduct with sample selection has been proposed. Different from the traditional attribute-reduction algorithm that only uses one measure, our algorithm is executed based on the multiple criteria, which include approximation quality and conditional entropy. Experimental results show that the reduct computed by our algorithm can not only increase approximation quality and preserve conditional entropy, but also provide better classification performance. Since we also applied boundary samples instead of the whole samples in the data, our algorithm needed to spend less time in finding reducts.

The following topics merit further investigations:

1. Only two measures have been used to design multiple criteria; some other measures, such as classification accuracy [57] and neighborhood discrimination index [12], will be further added into the construction of multiple criteria.
2. Multiple-criteria attribute reduction is realized by a neighborhood rough set; it can also be introduced into other rough-set models, such as a fuzzy rough set [19] and decision-theoretic rough set [58].
3. Attribute reduction can be considered as the first step of data processing, and classification performances in terms of different classifiers [59,60] based on our reducts will be further explored.

**Author Contributions:** project administration, X.Y.; software, P.W.; supervision, X.C.; writing—review and editing, Y.G.

**Funding:** This research was funded by the Natural Science Foundation of China (Nos. 61572242, 61502211, 61503160).

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

## References

1. Pawlak, Z. *Rough Sets: Theoretical Aspects of Reasoning about Data*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1991; ISBN 978-0792314721.
2. Pawlak, Z.; Skowron, A. Rough sets: Some extensions. *Inf. Sci.* **2007**, *177*, 28–40. [[CrossRef](#)]
3. Chen, H.M.; Li, T.R.; Luo, C.; Horng, S.J.; Wang, G. A decision-theoretic rough set approach for dynamic data mining. *IEEE Trans. Fuzzy Syst.* **2015**, *23*, 1–14. [[CrossRef](#)]
4. Kaneiwa, K.; Kudo, Y. A sequential pattern mining algorithm using rough set theory. *Int. J. Approx. Reason.* **2011**, *52*, 881–893. [[CrossRef](#)]
5. Hu, Q.H.; Yu, D.R.; Xie, Z.X.; Li, X. EROS: Ensemble rough subspaces. *Pattern Recognit.* **2007**, *40*, 3728–3739. [[CrossRef](#)]
6. Dowlatabadi, M.B.; Derhami, V.; Nezamabadi, P.H. Ensemble of filter-based rankers to guide an epsilon-greedy swarm optimizer for high-dimensional feature subset selection. *Information* **2017**, *8*, 152. [[CrossRef](#)]
7. Yao, Y.Y.; Zhao, Y. Attribute reduction in decision-theoretic rough set models. *Inf. Sci.* **2008**, *178*, 3356–3373. [[CrossRef](#)]
8. Hu, Q.H.; Yu, D.R.; Xie, Z.X. Neighborhood classifiers. *Expert Syst. Appl.* **2008**, *34*, 866–876. [[CrossRef](#)]
9. Dai, J.H.; Wang, W.T.; Xu, Q.; Tian, H. Uncertainty measurement for interval-valued decision systems based on extended conditional entropy. *Knowl.-Based Syst.* **2012**, *27*, 443–450. [[CrossRef](#)]
10. Dai, J.H.; Xu, Q.; Wang, W.T.; Tian, H. Conditional entropy for incomplete decision systems and its application in data mining. *Int. J. Gen. Syst.* **2012**, *41*, 713–728. [[CrossRef](#)]
11. Dai, J.H.; Wang, W.T.; Tian, H.W.; Liu, L. Attribute selection based on a new conditional entropy for incomplete decision systems. *Knowl.-Based Syst.* **2013**, *39*, 207–213. [[CrossRef](#)]

12. Wang, C.Z.; Hu, Q.H.; Wang, X.Z.; Chen, D.; Qian, Y.; Dong, Z. Feature selection based on neighborhood discrimination index. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 2986–2999. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Angiulli, F. Fast nearest neighbor condensation for large data sets classification. *IEEE Trans. Knowl. Data Eng.* **2007**, *19*, 1450–1464. [\[CrossRef\]](#)
14. Li, Y.H.; Maguire, L. Selecting critical patterns based on local geometrical and statistical information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1189–1201. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Nicolai, G.P.; Javier, P.R.; De, H.G.A. Oligois: Scalable instance selection for class-imbalanced data sets. *IEEE Trans. Cybern.* **2013**, *43*, 332–346. [\[CrossRef\]](#)
16. Lin, W.C.; Tsai, C.F.; Ke, S.W.; Hung, C.W. Learning to detect representative data for large scale instance selection. *J. Syst. Softw.* **2015**, *106*, 1–8. [\[CrossRef\]](#)
17. Zhai, J.H.; Wang, X.Z.; Pang, X.H. Voting-based instance selection from large data sets with mapreduce and random weight networks. *Inf. Sci.* **2016**, *23*, 1066–1077. [\[CrossRef\]](#)
18. Zhai, J.H.; Li, T.; Wang, X.Z. A cross-selection instance algorithm. *J. Intell. Fuzzy Syst.* **2016**, *3*, 717–728. [\[CrossRef\]](#)
19. Zhang, X.; Mei, C.L.; Chen, D.G.; Li, J. Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy. *Pattern Recognit.* **2016**, *56*, 1–15. [\[CrossRef\]](#)
20. Xu, S.P.; Yang, X.B.; Yu, H.L.; Yang, J.; Tsang, E.C. Multi-label learning with label-specific feature reduction. *Knowl.-Based Syst.* **2016**, *104*, 52–61. [\[CrossRef\]](#)
21. Yang, X.B.; Yao, Y.Y. Ensemble selector for attribute reduction. *Appl. Soft Comput.* **2018**, *70*, 1–11. [\[CrossRef\]](#)
22. Ju, H.R.; Yang, X.B.; Song, X.N.; Qi, Y. Dynamic updating multigranulation fuzzy rough set: Approximations and reducts. *Int. J. Mach. Learn. Cybern.* **2014**, *5*, 981–990. [\[CrossRef\]](#)
23. Yang, X.B.; Yu, D.J.; Yang, J.Y.; Wei, L. Dominance-based rough set approach to incomplete interval-valued information system. *Data Knowl. Eng.* **2009**, *68*, 1331–1347. [\[CrossRef\]](#)
24. Yao, Y.Y. Relational interpretations of neighborhood operators and rough set approximation operators. *Inf. Sci.* **1998**, *111*, 239–259. [\[CrossRef\]](#)
25. Yang, X.B.; Qian, Y.H.; Yang, J.Y. Hierarchical structures on multigranulation spaces. *J. Comput. Sci. Technol.* **2012**, *27*, 1169–1183. [\[CrossRef\]](#)
26. Yang, X.B.; Qi, Y.S.; Song, X.N.; Yang, J. Test cost sensitive multigranulation rough set: Model and minimal cost selection. *Inf. Sci.* **2013**, *250*, 184–199. [\[CrossRef\]](#)
27. Chen, D.G.; Wang, C.Z.; Hu, Q.H. A new approach to attribute reduction of consistent and inconsistent covering decision systems with covering rough sets. *Inf. Sci.* **2007**, *177*, 3500–3518. [\[CrossRef\]](#)
28. Hu, Q.H.; Pedrycz, W.; Yu, D.R.; Lang, J. Selecting discrete and continuous features based on neighborhood decision error minimization. *IEEE Trans. Syst. Man Cybern. B* **2010**, *40*, 137–150. [\[CrossRef\]](#)
29. Zhang, X.; Mei, C.L.; Chen, D.G.; Li, J. Multi-confidence rule acquisition and confidence-preserved attribute reduction in interval-valued decision systems. *Int. J. Approx. Reason.* **2014**, *55*, 1787–1804. [\[CrossRef\]](#)
30. Hu, Q.H.; Che, X.J.; Zhang, L.; Guo, M.; Yu, D. Rank entropy based decision trees for monotonic classification. *IEEE Trans. Knowl. Data Eng.* **2012**, *24*, 2052–2064. [\[CrossRef\]](#)
31. Liu, J.F.; Hu, Q.H.; Yu, D.R. A weighted rough set based method developed for class imbalance learning. *Inf. Sci.* **2008**, *178*, 1235–1256. [\[CrossRef\]](#)
32. Guo, G.D.; Wang, H.; Bell, D.; Bi, Y.; Greer, K. KNN model-based approach in classification. *Lect. Notes Comput. Sci.* **2003**, *2888*, 986–996. [\[CrossRef\]](#)
33. Li, S.Q.; Harner, E.J.; Adjeroh, D.A. Random knn feature selection—A fast and stable alternative to random forests. *BMC Bioinform.* **2011**, *12*, 450. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Sahigara, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Defining a novel  $k$ -nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. *J. Chem.* **2013**, *5*, 27–36. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Lin, G.P.; Liang, J.Y.; Qian, Y.H. Uncertainty measures for multigranulation approximation space. *Knowl.-Based Syst.* **2015**, *23*, 443–457. [\[CrossRef\]](#)
36. Li, M.M.; Zhang, X.Y. Information fusion in a multi-source incomplete information system based on information entropy. *Entropy* **2017**, *19*, 570. [\[CrossRef\]](#)

37. Karevan, Z.; Suykens, J.A.K. Transductive feature selection using clustering-based sample entropy for temperature prediction in weather forecasting. *Entropy* **2018**, *20*, 264. [[CrossRef](#)]
38. Ju, H.R.; Li, H.X.; Yang, X.B.; Zhou, X.; Huang, B. Cost-sensitive rough set: A multi-granulation approach. *Knowl.-Based Syst.* **2017**, *123*, 137–153. [[CrossRef](#)]
39. Dou, H.L.; Yang, X.B.; Song, X.N.; Yu, H.; Wu, W.Z.; Yang, J. Decision-theoretic rough set: A multicost strategy. *Knowl.-Based Syst.* **2016**, *91*, 71–83. [[CrossRef](#)]
40. Jia, X.Y.; Shang, L.; Zhou, B.; Yao, Y. Generalized attribute reduct in rough set theory. *Knowl.-Based Syst.* **2016**, *91*, 204–218. [[CrossRef](#)]
41. Li, H.X.; Zhou, X.Z. Risk decision making based on decision-theoretic rough set: A three-way view decision model. *Int. J. Comput. Intell. Syst.* **2011**, *4*, 1–11. [[CrossRef](#)]
42. Qian, Y.H.; Liang, J.Y.; Pedrycz, W.; Dang, C. Positive approximation: An accelerator for attribute reduction in rough set theory. *Artif. Intell.* **2010**, *174*, 597–618. [[CrossRef](#)]
43. Qian, Y.H.; Liang, J.Y.; Pedrycz, W.; Dang, C. An efficient accelerator for attribute reduction from incomplete data in rough set framework. *Pattern Recognit.* **2011**, *44*, 1658–1670. [[CrossRef](#)]
44. Jensen, R.; Shen, Q. Fuzzy-rough sets assisted attribute selection. *IEEE Trans. Fuzzy Syst.* **2007**, *15*, 73–89. [[CrossRef](#)]
45. Li, J.Z.; Yang, X.B.; Song X.N.; Li, J.; Wang, P.; Yu, D.J. Neighborhood attribute reduction: A multi-criterion approach. *Int. J. Mach. Learn. Cybern.* **2017**, 1–12. [[CrossRef](#)]
46. Dash, M.; Liu, H. Consistency-based search in feature selection. *Artif. Intell.* **2003**, *151*, 155–176. [[CrossRef](#)]
47. Hu, Q.H.; Pan, W.W.; Zhang, L.; Zhang, D.; Song, Y.; Guo, M.; Yu, D. Feature selection for monotonic classification. *IEEE Trans. Fuzzy Syst.* **2012**, *20*, 69–81. [[CrossRef](#)]
48. Wilson, D.R.; Martinez, T.R. Reduction techniques for instance-based learning algorithms. *Mach. Learn.* **2000**, *38*, 257–286.1007626913721. [[CrossRef](#)]
49. Brighton, H.; Mellish, C. Advances in instance selection for instance-based learning algorithms. *Data Min. Knowl. Discov.* **2002**, *6*, 153–172. [[CrossRef](#)]
50. Nikolaidis, K.; Goulermas, J.Y.; Wu, Q.H. A class boundary preserving algorithm for data condensation. *Pattern Recognit.* **2011**, *44*, 704–715. [[CrossRef](#)]
51. Aldahdooh, R.T.; Ashour, W. DIMK-means distance-based initialization method for *k*-means clustering algorithm. *Int. J. Intell. Syst. Appl.* **2013**, *5*, 41–51. [[CrossRef](#)]
52. Huang, K.Y. An enhanced classification method comprising a genetic algorithm, rough set theory and a modified PBMF-index function. *Appl. Soft. Comput.* **2012**, *12*, 46–63. [[CrossRef](#)]
53. Lingras, P.; Chen, M.; Miao, D. Qualitative and quantitative combinations of crisp and rough clustering schemes using dominance relations. *Int. J. Approx. Reason.* **2014**, *55*, 238–258. [[CrossRef](#)]
54. Yang, J.; Ma, Y.; Zhang, X.F.; Li, S.; Zhang, Y. An initialization method based on hybrid distance for *k*-means algorithm. *Neural Comput.* **2017**, *29*, 3094–3117. [[CrossRef](#)] [[PubMed](#)]
55. Vashist, R.; Garg, M.L. Rule generation based on reduct and core: A rough set approach. *Int. J. Comput. Appl.* **2011**, *29*, 1–5. [[CrossRef](#)]
56. Wang, G.Y.; Ma, X.A.; Yu, H. Monotonic uncertainty measures for attribute reduction in probabilistic rough set model. *Int. J. Approx. Reason.* **2015**, *59*, 41–67. [[CrossRef](#)]
57. Peng, H.C.; Long, F.H.; Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [[CrossRef](#)] [[PubMed](#)]
58. Azam, N.; Yao, J.T. Game-theoretic rough sets for recommender systems. *Knowl.-Based Syst.* **2014**, *72*, 96–107. [[CrossRef](#)]

59. Korytkowski, M.; Rutkowski, L.; Scherer, R. Fast image classification by boosting fuzzy classifiers. *Inf. Sci.* **2015**, *327*, 175–182. [[CrossRef](#)]
60. Tsang, E.C.C.; Hu, Q.H.; Chen, D.G. Feature and instance reduction for PNN classifiers based on fuzzy rough sets. *Int. J. Mach. Learn. Cybern.* **2016**, *7*, 1–11. [[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/>).