JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Ensemble of Classifiers Constructed on Class-Oriented Attribute Reduction

Min Li[*], Shaobo Deng[*], and Lei Wang[*]

Abstract

Many heuristic attribute reduction algorithms have been proposed to find a single reduct that functions as the entire set of original attributes without loss of classification capability; however, the proposed reducts are not always perfect for these multiclass datasets. In this study, based on a probabilistic rough set model, we propose the class-oriented attribute reduction (COAR) algorithm, which separately finds a reduct for each target class. Thus, there is a strong dependence between a reduct and its target class. Consequently, we propose a type of ensemble constructed on a group of classifiers based on class-oriented reducts with a customized weighted majority voting strategy. We evaluated the performance of our proposed algorithm based on five real multiclass datasets. Experimental results confirm the superiority of the proposed method in terms of four general evaluation metrics.

Keywords

Class-Oriented Attribute Reduction, Ensemble learning, Multiclass Datasets, Probabilistic Rough Sets

1. Introduction

In many cases, the classification and regression process can be better solved in the reduced attribute space compared with the original attribute space. An important topic in rough set theory [1] focuses on reducing the number of attributes in databases, which is called attribute reduction. It is an important pre-processing task in machine learning. The task of attribute reduction is to find a subset of the attributes that function as the entire set of available attributes without loss of classification capability. This subset is called a reduct. The merits of attribute reduction based on the rough set theory are reflected in maintaining the underlying semantics of the attributes without additional information for data analysis [2].

There are many types of attribute reductions in the field of rough sets, such as positive region reduction [3], assignment reduction [4,5], probabilistic distribution reduction [6], maximum probabilistic distribution reduction [7], variable precision reduction [8-10], M-reduction [11], and minimum cost attribute reduction [12,13]. All of these types of reducts can be sought using exhaustive algorithms and heuristic algorithms. Exhaustive algorithms, such as the discernibility matrix-based algorithms, generate all reducts of a dataset. However, generating all possible reducts from a dataset is a computationally intractable

---

problem because of the combinatorial explosion in the number of set attributes. For most applications, only one reduct is required, and the heuristic reduction algorithms appear far more attractive owing to their high efficiency. In most of these cases, the heuristic algorithms adopt a forward-greedy strategy to find a reduct on the basis of various measures of significance of attributes [14-16]. Thereafter, the found reduct is used to train a specific classifier for the classification task; however, a single reduct is not always perfect for these multiclass imbalanced datasets. For many multiclass imbalanced datasets, a single reduct usually includes attributes with poor classification accuracy for the minority classes. The heuristic attribute reduction algorithms evaluate a candidate attribute according to its significance measure oriented to the entire dataset, which results in a bias towards majority classes. Consequently, in this study, we separately seek a reduct for each target to maintain the classification performance of all classes for a dataset. Because there is a strong dependence between a reduct and its target class, the reduct can deliver a strong classification performance on its target class. Accordingly, we generate a base classifier on each class-oriented reduct and construct a type of ensemble of classifiers [17]. Previous studies show that a good ensemble depends on both the accuracy and diversity of the base classifiers [18]. Our proposed ensemble is diverse because it induces a group of classifiers based on class-oriented reducts; however, in many cases, the base classifier of our proposed ensemble has weak classification performance on objects in the dataset, except for objects of its target class. Put differently, the base classifiers of the proposed method are often weak for a whole dataset. To integrate results from a group of weak classifiers into one high-quality ensemble, a vote-weight matrix is designed. This allows the ensemble to achieve good performance. The main contributions of this study are summarized as follows:

- First, we propose the COAR algorithm based on a probabilistic rough model. The main merit of the COAR algorithm is that it can find a group of class-oriented reducts from a dataset without additional information.
- We propose a type of ensemble constructed on a group of classifiers based on class-oriented reducts. A customized and weighted majority voting strategy is designed. It promotes good performance of the proposed ensemble of classifiers.
- We conducted experiments using real multiclass datasets. Experimental results confirm the superiority of the proposed ensemble of classifiers compared with the classifier based on a single reduct.

The outline of this study is organized as follows: Section 2.1 presents preliminary concepts related to the rough set theory and the classical attribute reduction. Section 2.2 reviews the concepts of probabilistic distribution reduction and quick distribution reduction algorithm (Q-DRA). The concept of class-oriented probabilistic distribution reduction and the class-oriented attribute reduction (COAR) algorithm are described in Section 3.1. Section 3.2 presents the ensemble of classifiers constructed on COAR and introduces a type of weighted majority voting strategy. The empirical results and analysis are presented in Section 4. Conclusions and future research directions are discussed in Section 5.

2. Related Work

In this section, the basic concepts of rough set theory, such as the positive region reduct and the QuickReduct [19] algorithm are described in Section 2.1. The concept of probabilistic distribution reduction and Q-DRA is presented in Section 2.2.

2.1 Basic Concepts

In general, a dataset is described as a two-dimensional table, where each row expresses facts about an object and each column expresses a feature of that object. In rough set theory, such a two-dimensional table is called an information system. Formally, an information system is usually expressed in the form of $IS = (U, A, V, f_a)_{a \in A}$, where

$U = \{x_1, x_2, \cdots, x_n\}$ is a set of finite and non-empty objects; A is a non-empty finite set of attributes;
V is a set of values $= U_{a \in A} V_a$, where V_a is the domain of the attribute a;
f is an information function denoted by $f: U \times A \rightarrow V$, which specifies the attribute values of x from U.

With attributes $B \subseteq A$, $Ind(B) = \{(x, y) \in U \times U | \forall a \in B, f_a(x) = f_a(y)\}$ is an equivalence relation, where $f_a(x)$ and $f_a(y)$ denote the values of objects x and y of the attribute a, respectively. $Ind(B)$ divides U into a family of disjoint granules, which are denoted by $U/B = \{X_1, X_2, \cdots, X_k\}$, where X_i ($i = 1, 2, \cdots, k$) is an equivalence granule induced by $Ind(B)$.

Let $X \subseteq U$ and $B \subseteq A$, X can be characterized using only the information contained in the attributes B using two unions of elementary sets $\underline{B}(X) = \{x \in U | [x]_B \subseteq X\}$ and $\overline{B}(X) = \{x \in U | [x]_B \cap X \neq \emptyset\}$, where $[x]_B = \{y \in U | (x, y) \in Ind(B)\}$, $\underline{B}(X)$ and $\overline{B}(X)$ are called B-lower and B-upper approximations, respectively. The set $BN_B(X) = \overline{B}(X) - \underline{B}(X)$ is known as the boundary region of X. X is termed as definable if the boundary region $BN_B(X) = \emptyset$. Otherwise, X is said to be rough.

In supervised machine learning, the class label of each object is known. In such a situation, an information table can be referred to as a decision table, which is represented as $DT = (U, A = C \cup D, V, f)$, $C \cap D = \emptyset$, where C is known as the condition attribute, and D is called the decision attribute. For $B \subseteq C$, $POS_B(D) = \cup_{X \in U/D} \underline{B}(X)$ is known as the positive region of decision D with respect to B.

For a decision table $DT = (U, C \cup D, V, f)$, an attribute $a \in B \subseteq C$ is relatively dispensable in B if $POS_{B-\{a\}}(D) = POS_B(D)$; otherwise, a is said to be relatively indispensable in B. If every attribute in B is relatively indispensable, we say that attributes B are relatively independent in DT. An attribute set $B \subseteq C$ is a relative positive region reduct if $POS_B(D) = POS_C(D)$ and B is relatively independent in DT.

A relative positive region reduct can maintain the discernibility power of those objects in the original positive region. Generally, only one relative reduct is required. One of the most widely applied measures of attribute significance is defined as follows [20]:

$$SIG(a, B, D) = \gamma_{B \cup \{a\}}(D) - \gamma_B(D); \tag{1}$$

$$\gamma_B(D) = \frac{|POS_B(D)|}{|U|} = \frac{\sum_{Y \in U/B \wedge |Y/D|=1} |Y|}{|U|}. \tag{2}$$

Formula (1) denotes the significance of a introduced in B. Formula (2) defines the degree of dependency of D by attributes B. Clearly, $0 \leq \gamma_B(D) \leq 1$. Based on the above definitions, the QuickReduct [19] algorithm is typical in heuristic algorithms that use a greedy search strategy to seek a relative positive region reduct. The pseudo-code of the QuickReduct is as follows:

Algorithm 1. QuickReduct algorithm

Input: Decision table $DT = (U, C \cup D, V, f)$
Output: One relative position region reduction of DT.
   1 $red \leftarrow \{\}$
   2 do
   3    $T \leftarrow red$
   4   **foreach** $a \in C - red$
   5   **if** $\gamma_{red \cup \{a\}}(D) > \gamma_T(D)$
   6        $T \leftarrow red \cup \{a\}$
   7    $red \leftarrow T$
   8 **until** $\gamma_{red}(D) = \gamma_C(D)$
   9 $return\ red$

$\gamma_B(D)$ cannot maintain the discernibility power of those objects in the boundary region; consequently, the authors [7] introduced a concept of distribution positive dependency degree that can retain probabilistic rough membership for the very object. In the next subsection, the most relevant notions and algorithm of probabilistic distribution reduction are briefly reviewed.

2.2 Probabilistic Distribution Reduction

Precisely, the rough membership of x belonging to set X is expressed as follows:

$$\mu_X^B(x) = p(X|[x]_B) = \frac{|[x]_B \cap X|}{|[x]_B|} \tag{3}$$

For a decision table, $DT = (U, C \cup D, V, f)$, $B \subseteq C \cup D$, and $X \subseteq U(X \neq \emptyset)$. The distribution approximation set of X based on the attribute set B is defined as

$$\tilde{B}^d(X) = \left\{ \frac{\mu_X^B(x)}{x} | x \in X \right\}, \tag{4}$$

where symbol "d" means the probabilistic distribution approximation. Note that the classical rough set uses a pair of lower approximation $\underline{B}(X)$ and upper approximation $\overline{B}(X)$ to characterize X, whereas X can be precisely described by a probabilistic distribution approximation set $\tilde{B}^d(X)$. The distribution approximation quality of X by the attributes B is defined as

$$r_B^d(X) = \frac{|\tilde{B}^d(X)|}{|X|} = \frac{\sum_{x \in X} \mu_X^B(x)}{|X|}. \tag{5}$$

$r_B^d(X)$ can also be expressed as follows:

$$r_B^d(X) = \frac{|\underline{B}(X)| + \sum_{x \in X - \underline{B}(X)} \mu_X^B(x)}{|X|}, \tag{6}$$

$$r_B^d(X) = \frac{1}{|X|} \sum_{Y \in U/B} \frac{|Y \cap X|^2}{|Y|}. \tag{7}$$

According to Formula (6), we can compare the discernibility power of attributes P and B in the case of $\underline{P}(X) = \underline{B}(X)$. In addition, $r_B^d(X) \geq r_{B'}^d(X)$ holds, if $B' \subseteq$ B. The positive rough membership value of an object $x \in U$ by the attributes C is defined by:

$$\mu_{pos}^C(x) = p([x]_D|[x]_C) = \frac{|[x]_{C \cup D}|}{|[x]_C|}. \tag{8}$$

The distribution positive region of U based on attributes B is defined as follows:

$$POS_B^d(D) = \bigcup_{X \in U/D} \tilde{B}^d(X) = \left\{ \frac{\mu_{pos}^B(x)}{x} \middle| x \in U \right\}. \tag{9}$$

The distribution positive dependency degree of U based on the attributes B is defined by:

$$\gamma_B^d(U) = \frac{|POS_B^d(D)|}{|U|} = \frac{1}{|U|} \sum_{Y \in U/B} \sum_{X \in U/D} \frac{|Y \cap X|^2}{|Y|}. \tag{10}$$

Clearly, $\gamma_B^d(U) \geq \gamma_{B'}^d(U)$ holds, if $B' \subseteq$ B. The distribution significance of the attribute a with respect to B is defined by

$$SIG^d(a, B, D) = \gamma_{B \cup \{a\}}^d(D) - \gamma_B^d(D). \tag{11}$$

Based on Formula (11), a hill-climbing search algorithm Q-DRA [7, 21] is listed below for finding a probabilistic distribution reduct.

Algorithm 2. Q-DRA

Input: Decision table $DT = (U, C \cup D, V, f)$
Output: One probabilistic distribution reduction of DT.

```
 1  red ← {}
 2  do
 3  │  T ← red
 4  │  flag ← true
 5  │  foreach a ∈ C − red
 6  │  │  if γ^d_{red∪{a}}(D) > γ^d_T(D)
 7  │  └──T ← red ∪ {a}
 8  │  if T ≠ red
 9  │  │   red ← T
10  │  └─flag ← false
11  └ until flag = true
12  return red
```

---

3. Ensemble of Classifiers Constructed on Class-Oriented Reducts

In this section, we present a novel ensemble constructed on a group of class-oriented probabilistic distribution reducts. In Section 3.1, we present the COAR algorithm, which is able to obtain a probabilistic distribution reduct for each decision class of a dataset. In Section 3.2, a type of ensemble with a

customized weighted majority voting strategy is designed.

3.1 Class-Oriented Probabilistic Distribution Reduction

In many cases, only a probabilistic distribution reduct (or relative position region reduct) is needed because it can preserve distribution (or position region) discernibility power and the consequent classification performance. However, this does not mean that a single reduct is always well suited for each class in a dataset, especially for datasets with class imbalance and data drift. For many multiclass datasets, a single reduct frequently includes some attributes with poor classification performance for minority classes because of the global view of the search process of the heuristic attribute reduction algorithms. Consequently, we tend to separately seek one reduct for each target class to maintain classification performance for all classes of a dataset.

The distribution approximation quality is an effective uncertainty measurement, which can be used to calculate the probabilistic distribution discernibility power of attributes B with respect to the decision class $X(X \in U/D)$. The distribution significance of attributes a relative to attributes B with respect to decision class X is defined as follows:

$$SIG^d(a, B, X) = r_{B \cup \{a\}}^d(X) - r_B^d(X) . \tag{12}$$

The COAR algorithm list below is used for obtain a probabilistic distribution reduct for each decision class of a dataset.

Algorithm 3. COAR

Input: Decision table $DT = (U, C \cup D, V, f)$
Output: a probabilistic distribution reduction for each $X \in U/D$
    1 $redList \leftarrow \{\}$
    2 foreach $X \in U/D$
    3    $red \leftarrow \{\}$
    4   do
    5      $T \leftarrow red$
    6      $flag \leftarrow true$
    7   **foreach** $a \in C - red$
    8   **if** $r_{red \cup \{a\}}^d(X) > r_T^d(X)$
    9        $T \leftarrow red \cup \{a\}$
  10   **if** $T \neq red$
  11      $red \leftarrow T$
  12      $flag \leftarrow false$
  13   **until** $flag = true$
  14 Add red into $redList$
  15 return $redList$

Note that the COAR algorithm returns a *redList* that includes a group of class-oriented reducts. Given a dataset DT containing n samples with m features, the time complexity for seeking a class-oriented probabilistic distribution reduct might need to perform $(m^2 + m)/2$ evaluations of the distribution approximation quality for the worst-case. If a radix sorting technique is adopted, the time complexity of

the COAR algorithm is $O(knm^2)$ for the worst-case, where k is the number of classes.

3.2 Construction of Ensemble of Classifiers

COAR seeks a reduct for each target class separately; therefore, it can find a group of diverse reducts for a multiclass dataset. Each reduct is selected for a target class, which has the benefit of separating the class from others. Put differently, whether an object belongs to a specific class can be ascertained by using the corresponding reduct. We generate a base classifier on each reduct and obtain a group of base classifiers. Thereafter, an ensemble of classifiers can be constructed. Previous studies show that a good ensemble depends on both the accuracy and diversity of the base classifiers. For some datasets (especially for datasets with many classes), we observed that a single reduct in many instances includes weak classification performance on objects in a dataset except for its target class. Put differently, these base classifiers are usually weak classifiers on the entire dataset. To obtain the results from a group of weak classifiers into one high-quality ensemble classifier, a vote-weight matrix VW is designed as Formula (13).

$$VW(i,j) = \begin{bmatrix} k & 1 & \cdots & 1 \\ 1 & k & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & \cdots & k \end{bmatrix} \tag{13}$$

VW is a matrix with k rows and k columns, where k is the number of classes of a dataset. All diagonal entries are set to, k and other entries are set to 1. The ith row of VW is the weight vector of the ith base classifier (is constructed on a reduct oriented ith class), which means the vote-weight is $VW(i,j)$ if the ith base classifier categorizes an object into the jth class. Therefore, we design a type of weighted majority voting strategy to classify an object. For an object, let $C(i,j) = 1$ denote the ith base classifier that categorizes the object into the jth class; otherwise, let $C(i,j) = 0$. Thereafter, the score of an object is classified into the jth class by an ensemble of classifiers and is computed as follows:

$$Vote[j] = \sum_{i=1}^{k} C(i,j) \times VW(i,j) \tag{14}$$

Finally, the prediction class of an object by the ensemble of classifiers corresponds to the maximum value of *Vote*. The reasons for setting diagonal entries of the matrix k are twofold. First, the ith base classifier categorizes an object into the ith class, usually with a high degree of credibility. Second, the result cannot be turned down if the ith base classifier categorizes an object into the ith class except that the $jth(j \neq i)$ base classifier categorizes this object into the jth class.

4. Experimental Results and Analyses

Our motivation for seeking class-oriented reducts is that a single reduct is not always well suited for each class of datasets for prediction purposes. Our proposed ensemble is constructed on a group of classifiers based on class-oriented reducts. To demonstrate the effectiveness of the proposed ensemble of classifiers, we compare it with the classifier constructed on a single reduct of the dataset. It has been

proven that the Q-DRA has clear advantages over the QuickReduct (Algorithm 1). Hence, we focus on the comparison with the attribute selection technique of Q-DRA. In the experiments, two popular classification algorithms, C4.5 [22] and RBF-SVM (support vector machine with radial basis function kernel) [23] are introduced to induce base classifiers.

4.1 Benchmark Datasets

We experimented with five different types of real-life multiclass datasets: GLASS, ZOO, DERMATOLOGY, SOYBEAN-LARGE, and VOWEL, which were obtained from the UCI Machine Learning Repository.

Table 1 presents the reducts found by the Q-DRA and some examples of reducts found by the COAR for a special dataset. The integers listed in the second and fourth columns in Table 1 represent the corresponding condition attributes in each specific dataset, i.e., 1 denotes the first condition attribute, 2 denotes the second condition attribute, and so on. As shown in Table 1, the number of attributes in the reducts found by the Q-DRA is significantly lower than the number of the original attributes. For each dataset, the COAR generates a group of class-oriented reducts, and we present the average size (number of attributes) of the reducts. Note that the average size of the reducts found by COAR is substantially smaller than the size of the corresponding reducts found by Q-DRA. For example, in SOYBEAN-LARGE, the average size of the reducts found by the COAR is 3.5, compared with 13 that is the size of the reduct found by Q-DRA. This result is expected because the reduct found by the COAR only needs to discriminate a specific target class, whereas the reducts found by Q-DRA need to discriminate all classes.

Table 1. Comparison of attribute sequences of reducts

Dataset	Q-DRA		COAR	
	Reduct	Size	Some reducts with different size	Average size
GLASS	2,4,5,3,10,8,7	7	8,5,6,2 9,6,3,4,2 2,4,5,6,3,7	5.3
ZOO	13,4,6,8, 3	5	2 12,3 11,13,3,6,8	2.6
DERMATOLOGY	22,33,15,34,4,1,9	7	31,1 22,20,1 28,34,16,14,1	3.2
SOYBEAN-LARGE	29,22,15,1,4,3,18, 7,10,6,9,5,16	13	6 29,21,13 15,30,35,4	3.5
VOWEL	2,1,8,10,4,3	6	2,1,4 2,1,6,3 9,2,8,10,1	4.5

4.2 Evaluation Metrics

In our experiments, we examined the effect of four general indexes: overall accuracy, precision, recall, and f-measure [24]. All of these indexes are derived from the confusion matrix with the concepts of TP

(true positive), FP (false positive), TN (true negative), and FN (false negative). Let n denote the total number of instances. The overall accuracy is the simplest measure, and is defined as follows:

$$overall\ accuracy = \frac{\sum_{i=1}^{S} TP_i}{n}. \tag{15}$$

For the ith target class, the precision is defined as Formula (16):

$$precision_i = \frac{TP_i}{TP_i + FP_i}. \tag{16}$$

The recall of the ith target class is defined by Formula (17):

$$recall_i = \frac{TP_i}{TP_i + FN_i}. \tag{17}$$

The f-measure is an index based on a balance between precision and recall, which is defined by Formula (18).

$$f - measue_i = \frac{2 * precision_i * recall_i}{precision_i + recall_i}. \tag{18}$$

4.3 Classification Performance

The main objective of this section is to illustrate the advantage of the proposed ensemble constructed on a group of classifiers based on class-oriented attribute subsets compared with the classifiers based on a single attribute reduction set (Q-DRA). We adopted a 10-fold cross-validation method to analyze the classification results.
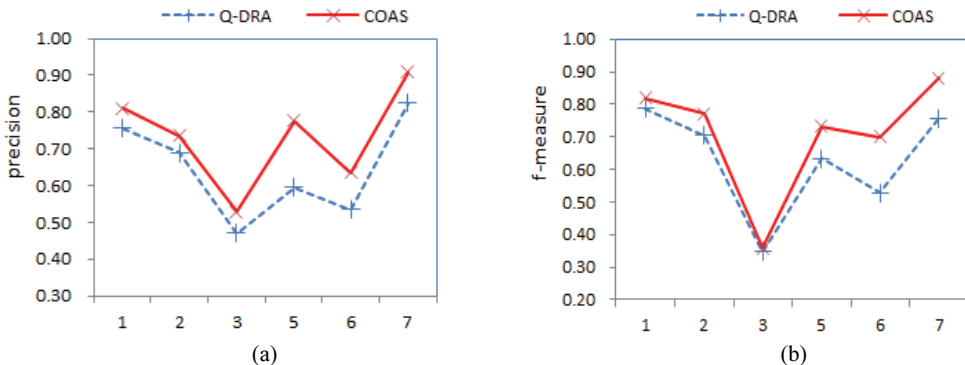
Fig. 1. Comparison of precision (a) and f-measure (b) using the C4.5 classifier for the GLASS dataset.

GLASS: The GLASS dataset has six decision classes. A scatter plot in Fig. 1 provides an intuitive comparison of precision and f-measure for every class using two methods. From Fig. 1, one can see that the COAR obtains all superior precisions and f-measures. For a quantitative analysis, Table 2 reports the precision, recall, and f-measure for every class of GLASS obtained on 10 runs using the C4.5 classifier. The direct comparison of evaluation metrics can be obtained by checking the average values in the last

row of Table 2. From the average values of the evaluation metrics, we notice the COAR obtains the best precision, recall, and f-measure. Further comparing the classification performance for each class of GLASS, the COAR obtains 6 times higher precision, 5 times higher recall, and 6 times higher f-measure. We notice that the Q-DRA obtains a slightly better recall index for only one class (class 3); however, it obtains relatively lower precision and f-measure for this class. In addition, Table 2 lists the precision, recall, and f-measure for every class of GLASS using the RBF-SVM classifier. A similar comparison result can be drawn: the COAR is obviously superior to Q-DRA when using the RBF-SVM classifier.

Table 2. Classification performances of two compared algorithms (GLASS dataset)

| Algorithm	Class name	Q-DRA			COAR		
		Precision	Recall	f-measure	Precision	Recall	f-measure
C4.5	1	0.756	0.821	0.787	0.81	0.826	0.818
	2	0.689	0.722	0.705	0.735	0.812	0.771
	3	0.47	0.276	0.348	0.529	0.271	0.358
	4	0.595	0.677	0.633	0.776	0.692	0.732
	5	0.534	0.522	0.528	0.636	0.778	0.700
	6	0.825	0.7	0.757	0.908	0.855	0.881
	Mean	0.645	0.620	0.626	0.732	0.706	0.710
RBF-SVM	1	0.699	0.854	0.769	0.757	0.88	0.814
	2	0.713	0.799	0.754	0.712	0.841	0.771
	3	0.571	0.071	0.126	0.98	0.282	0.438
	4	0.888	0.608	0.721	0.862	0.623	0.723
	5	0.769	0.556	0.645	0.879	0.567	0.689
	6	0.915	0.814	0.861	0.978	0.766	0.859
	Mean	0.759	0.617	0.646	0.861	0.660	0.716
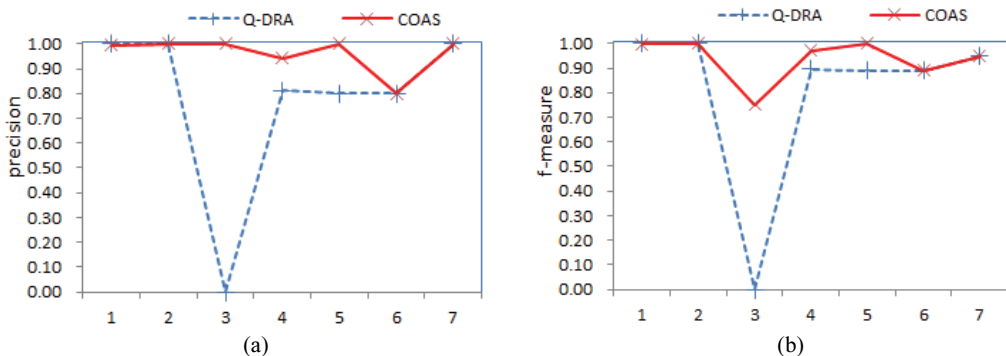
Fig. 2. Comparison of precision (a) and f-measure (b) using the C4.5 classifier for the ZOO dataset.

ZOO: The ZOO dataset contains seven decision classes. A scatter plot in Fig. 2 provides an intuitive comparison of precision and f-measure for every class using two methods. From Fig. 2, we can see that the COAR obtains almost all superior precisions and f-measures. For a quantitative analysis, Table 3 lists the precision, recall, and f-measure for every class of ZOO obtained on 10 runs using the C4.5 classifier. From the average values in the last row in Table 3, we find that the COAR obtains the best average values of precision, recall, and f-measure. It is evident from Table 3 that out of 7 classes, the COAR obtains 6

times (includes 2 times identical) higher precision, 7 times higher recall (includes 3 times identical), and 6 times higher f-measure. We notice that the Q-DRA obtains 0 values of precision, recall, and f-measure for class 3. In comparison, the COAR obtains the precision, recall, and f-measure values of 1.0, 0.6, and 0.75, respectively. It was observed that class 3 in ZOO is a minority class because it only includes five instances. This result shows that the COAR is better for helping in the classification of minority classes, which is crucial for the classification of imbalanced datasets. In addition, Table 3 lists the precision, recall, and f-measure for every class of ZOO using the RBF-SVM classifier. A similar comparison result can be drawn: the COAR is clearly superior to Q-DRA when using the RBF-SVM classifier.

Table 3. Classification performances of two compared algorithms (ZOO dataset)

Algorithm	Class name	Q-DRA			COAR		
		Precision	Recall	f-measure	Precision	Recall	f-measure
C4.5	1	1.000	1.000	1.000	0.995	1.000	0.998
	2	1.000	1.000	1.000	1.000	1.000	1.000
	3	0.000	0.000	0.000	1.000	0.600	0.750
	4	0.813	1.000	0.897	0.942	1.000	0.970
	5	0.800	1.000	0.889	1.000	1.000	1.000
	6	0.800	1.000	0.889	0.800	1.000	0.889
	Mean	1.000	0.900	0.947	1.000	0.900	0.947
RBF-SVM	1	0.983	0.971	0.977	0.993	1.000	0.996
	2	1.000	1.000	1.000	1.000	1.000	1.000
	3	0.000	0.000	0.000	0.000	0.000	0.000
	4	0.850	1.000	0.919	0.818	1.000	0.900
	5	0.656	1.000	0.792	0.597	1.000	0.748
	6	1.000	1.000	1.000	1.000	1.000	1.000
	Mean	0.811	0.900	0.853	0.989	0.900	0.942
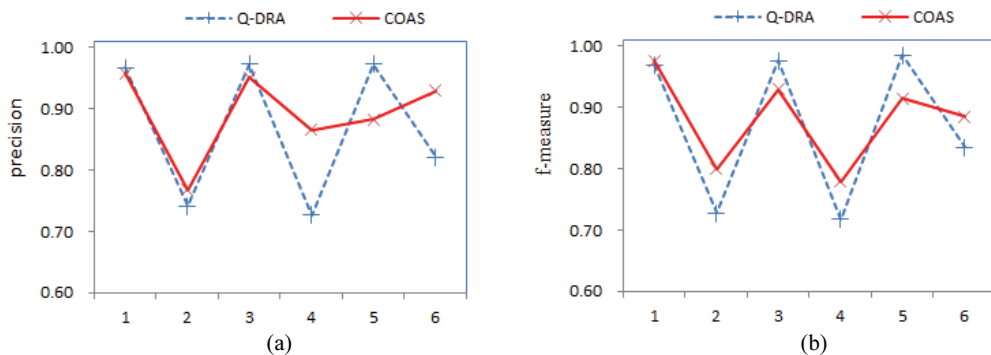
Fig. 3. Comparison of precision (a) and f-measure (b) using the C4.5 classifier for the DERMATOLOGY dataset.

DERMATOLOGY: The DERMATOLOGY dataset contains six decision classes. A scatter plot in Fig. 3 presents an intuitive comparison of precision and f-measure for every class using two methods. From Fig. 3, we can see that the COAR obtains slightly lower values of f-measure for class 3 and class 5; however, the COAR obtains a higher f-measure for the other four classes. Furthermore, we notice that

the COAR obtains higher f-measure values for the smallest class 6 (includes 20 instances) and the second smallest class 4 (includes 49 instances) in DERMATOLOGY. This result again shows that the COAR is better for helping in the classification of minority classes. For a quantitative analysis, Table 4 lists the precision, recall, and f-measure for every class of DERMATOLOGY obtained on 10 runs using the C4.5 classifier. From the average values in the last row in Table 4, we notice that the COAR again obtains the best average values of precision, recall, and f-measure. It is evident from Table 4 that out of 6 classes, the COAR obtains 3 times higher precision, 3 times higher recall, and 4 times higher f-measure. In addition, Table 4 lists the precision, recall, and f-measure for every class of DERMATOLOGY using the RBF-SVM classifier. It can also be seen that the COAR yielded better results compared with the Q-DRA for every evaluation metric when using the RBF-SVM classifier.

Table 4. Classification performances of two compared algorithms (DERMATOLOGY dataset)

Algorithm	Class name	Q-DRA			COAR		
		Precision	Recall	f-measure	Precision	Recall	f-measure
C4.5	1	0.966	0.971	0.969	0.957	0.996	0.976
	2	0.741	0.716	0.728	0.767	0.834	0.799
	3	0.972	0.979	0.976	0.951	0.908	0.929
	4	0.726	0.708	0.717	0.865	0.708	0.779
	5	0.972	0.998	0.985	0.882	0.948	0.914
	6	0.821	0.85	0.835	0.929	0.845	0.885
	Mean	0.866	0.870	0.868	0.892	0.873	0.880
RBF-SVM	1	0.899	0.959	0.928	0.869	0.997	0.929
	2	0.585	0.625	0.604	0.699	0.616	0.655
	3	0.984	0.961	0.973	0.915	0.929	0.922
	4	0.596	0.606	0.601	0.736	0.524	0.613
	5	0.973	0.765	0.857	0.906	0.979	0.941
	6	0.863	0.885	0.874	0.938	0.915	0.927
	Mean	0.817	0.800	0.806	0.844	0.827	0.831
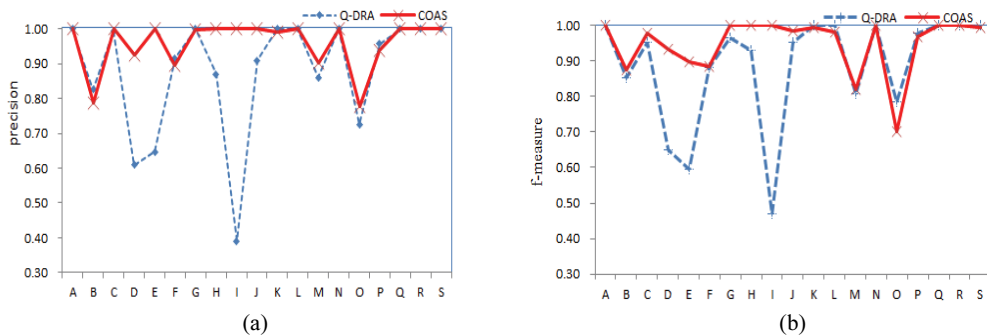
Fig. 4. Comparison of precision (a) and f-measure (b) using the C4.5 classifier for the SOYBEAN-LARGE dataset.
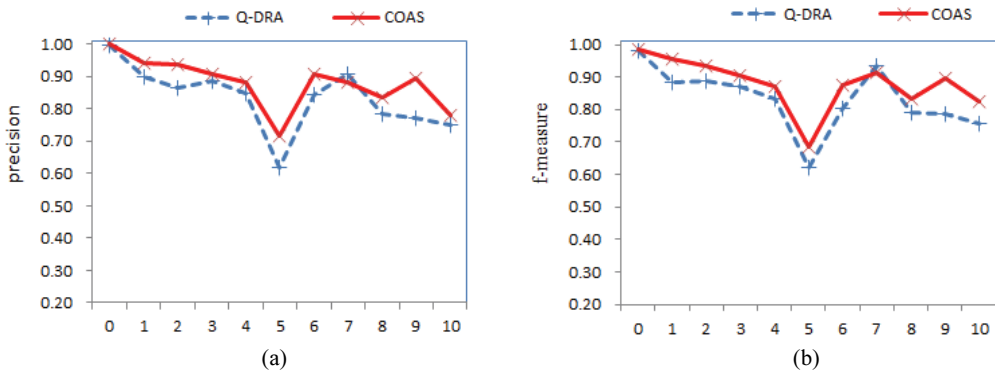
SOYBEAN-LARGE: The SOYBEAN-LARGE dataset contains 19 decision classes. A scatter plot in Fig. 4 presents an intuitive comparison of precision and f-measure for every class using two methods. It is evident from Fig. 4 that the COAR obtains almost all superior precisions and f-measures. From Fig. 4,

Table 5. Classification performances of two compared algorithms (SOYBEAN-LARGE dataset)

Algorithm	Class name	Q-DRA			COAR		
		Precision	Recall	f-measure	Precision	Recall	f-measure
C4.5	A	1.000	1.000	1.000	1.000	1.000	1.000
	B	0.826	0.879	0.852	0.787	0.982	0.874
	C	0.993	0.914	0.951	1.000	0.952	0.976
	D	0.609	0.700	0.651	0.926	0.940	0.933
	E	0.647	0.550	0.595	1.000	0.815	0.898
	F	0.916	0.846	0.880	0.896	0.870	0.883
	G	1.000	0.932	0.965	0.998	1.000	0.999
	H	0.870	1.000	0.930	1.000	1.000	1.000
	I	0.390	0.586	0.469	1.000	1.000	1.000
	J	0.909	1.000	0.952	1.000	0.967	0.983
	K	1.000	1.000	1.000	0.990	1.000	0.995
	L	1.000	1.000	1.000	1.000	0.965	0.982
	M	0.861	0.760	0.807	0.902	0.752	0.820
	N	1.000	1.000	1.000	1.000	1.000	1.000
	O	0.726	0.850	0.783	0.779	0.635	0.700
	P	0.958	0.997	0.977	0.940	1.000	0.969
	Q	1.000	1.000	1.000	1.000	1.000	1.000
	R	1.000	1.000	1.000	1.000	1.000	1.000
	S	1.000	1.000	1.000	1.000	0.990	0.995
	Mean	0.879	0.895	0.885	0.959	0.940	0.948
RBF-SVM	A	1.000	0.675	0.806	1.000	0.975	0.987
	B	0.713	0.968	0.821	0.693	0.971	0.809
	C	0.982	0.880	0.928	1.000	0.927	0.962
	D	0.727	0.785	0.755	0.927	0.955	0.941
	E	0.803	0.630	0.706	1.000	0.720	0.837
	F	0.763	0.855	0.807	0.790	0.849	0.819
	G	0.939	0.943	0.941	0.987	1.000	0.993
	H	0.892	0.995	0.941	1.000	1.000	1.000
	I	0.864	0.271	0.413	1.000	0.929	0.963
	J	1.000	0.867	0.929	1.000	0.867	0.929
	K	0.957	1.000	0.978	1.000	1.000	1.000
	L	0.952	1.000	0.976	1.000	0.815	0.898
	M	0.851	0.742	0.793	0.830	0.719	0.770
	N	1.000	0.713	0.832	1.000	0.838	0.912
	O	0.928	0.640	0.757	0.982	0.270	0.424
	P	0.984	0.993	0.989	0.938	1.000	0.968
	Q	1.000	0.850	0.919	1.000	0.990	0.995
	R	0.886	0.855	0.870	1.000	0.690	0.817
	S	1.000	0.955	0.977	1.000	1.000	1.000
	Mean	0.907	0.822	0.849	0.955	0.869	0.896

we notice that the COAR clearly obtains higher values of precision and f-measure for classes D, E, H, I, and J. Note that class D (includes 20 instances), class E (includes 20 instances), class H (includes 20 instances), class I (includes 14 instances), and class J (includes 15 instances) are all minorities. This result once again shows that the COAR is better for helping in the classification of minority classes. For a quantitative analysis, Table 5 lists the precision, recall, and f-measure for every class of the SOYBEAN-

LARGE obtained on 10 runs using the C4.5 classifier. According to the average values in the last row in Table 5, the COAR obtains clearly higher average values of precision, recall, and f-measure. From Table 5, it can be seen that the COAR obtains 14 times (including 8 identical times) higher precision, 14 times (includes 8 identical) higher recall, and 14 times (including 7 times identical) higher f-measure. In addition, Table 5 lists the precision, recall, and f-measure for every class of the SOYBEAN-LARGE using the RBF-SVM classifier. It can also be seen that the COAR yielded better results compared with the Q-DRA for every evaluation metric when using the RBF-SVM classifier.

(a) (b)

Fig. 5. Comparison of precision (a) and f-measure (b) using the C4.5 classifier for the VOWEL dataset.

VOWEL: The VOWEL dataset is a strict balanced dataset with 11 decision classes. A scatter plot in Fig. 5 provides an intuitive comparison of precision and f-measure for every class using two methods. From this figure, one can see that the COAR obtains almost all superior precisions and f-measures because the precision curve and f-measure curve of the COAR are above the precision curve and f-measure curve of the Q-DRA in most cases. For a quantitative analysis, Table 6 lists the precision, recall, and f-measure for every class of the VOWEL obtained on 10 runs using the C4.5 classifier. According to the average values in the last row in Table 6, the COAR obtains clearly higher average values of precision, recall, and f-measure. Further comparing the classification performance for each class of the VOWEL, the COAR obtains 10 times higher precision, 10 times higher recall, and 10 times higher f-measure. The Q-DRA obtains slightly better precision, recall, and f-measure for only one class (class 7). In addition, Table 6 lists the precision, recall, and f-measure for every class of VOWEL using the RBF-SVM classifier. It can also be seen that the COAR yielded better results compared with the Q-DRA for every evaluation metric when using the RBF-SVM classifier.

From Figs. 1–5, it is noticeable that the scatter plots of the Q-DRA show clear fluctuation because the Q-DRA obtains lower classification performance for some minority classes. These results indicate that a single reduct cannot maintain the classification performance of datasets in some cases, especially for those multiclass imbalanced datasets. In contrast, the COAR improves classification performance for minority classes, and concurrently maintains or improves classification performance for other classes. In addition, the experiments on the VOWEL dataset show that the COAR can also obtain superior classification performance for a multiclass balanced dataset compared with the Q-DRA. This confirms that a group of class-oriented reducts obtained by the COAR are beneficial for improving classification performance. Table 7 respectively lists the overall accuracy using the C4.5 and RBF-SVM. It is clear that the COAR obtains superior overall accuracy for every dataset.

Table 6. Classification performances of two compared algorithms (VOWEL dataset)

Algorithm	Class name	Q-DRA			COAR		
		Precision	Recall	f-measure	Precision	Recall	f-measure
C4.5	1	0.998	0.965	0.981	1.000	0.969	0.984
	2	0.897	0.873	0.885	0.941	0.967	0.954
	3	0.863	0.917	0.889	0.937	0.929	0.933
	4	0.885	0.863	0.873	0.906	0.908	0.907
	5	0.846	0.825	0.835	0.884	0.856	0.870
	6	0.617	0.625	0.621	0.715	0.654	0.683
	7	0.844	0.765	0.802	0.908	0.848	0.877
	8	0.908	0.963	0.934	0.883	0.946	0.913
	9	0.786	0.796	0.791	0.834	0.835	0.835
	10	0.772	0.804	0.788	0.896	0.898	0.897
	Mean	0.749	0.760	0.755	0.780	0.873	0.824
RBF-SVM	1	0.960	0.998	0.979	0.992	1.000	0.996
	2	0.949	0.967	0.958	0.970	0.994	0.981
	3	0.962	0.946	0.954	0.998	0.985	0.992
	4	0.956	0.942	0.949	0.974	0.942	0.958
	5	0.968	0.954	0.961	0.886	0.921	0.903
	6	0.786	0.842	0.813	0.812	0.844	0.827
	7	0.958	0.852	0.902	0.974	0.850	0.908
	8	0.953	0.979	0.966	0.931	0.954	0.942
	9	0.853	0.981	0.913	0.913	0.913	0.913
	10	0.945	0.890	0.916	0.956	0.985	0.970
	Mean	0.920	0.835	0.876	0.867	0.871	0.869

Table 7. Overall classification accuracy by C4.5 with Q-DRA and COAR

Data set	C4.5		RBF-SVM	
	Q-DRA	COAR	Q-DRA	COAR
GLASS	0.7051	0.7706	0.7393	0.7743
ZOO	0.9406	0.9703	0.9287	0.9406
DERMATOLOGY	0.8923	0.8984	0.8249	0.8500
SOYBEAN-LARGE	0.8859	0.9240	0.8630	0.8914
VOWEL	0.8322	0.8803	0.9259	0.9326

5. Conclusions

Rough set theory has been widely used as a tool of attribute reduction with significant success. Many rough-set-theory-based heuristic algorithms focus on finding a reduct that can provide the same information for classification purposes as the entire set of available attributes. However, for multiclass datasets, a single reduct (oriented whole dataset) often biases towards the majority classes. In addition, a single reduct is not always well suited for the purposes of prediction, even for balanced multiclass datasets. In this study, we propose a COAR algorithm that seeks one reduct for each target class separately so that there is a strong dependence between a reduct and its target class. Consequently, we propose a type of ensemble constructed on a group of classifiers based on class-oriented reducts with a customized

weighted majority voting strategy. We conducted experiments with five real multiclass datasets. The experimental results confirm the superiority of the COAR algorithm in terms of four general evaluation metrics.

There are several activities for future studies. First, in a situation where people are only concerned about determining whether an object belongs to a specific target class, class-oriented attribute-reduction algorithms may bring more benefits. Second, instead of finding one reduct for each class, we can find several different reducts for each class to strengthen the diversity of the ensemble. Finally, the voting strategy of our proposed ensemble can be further optimized for a concrete dataset.

Acknowledgement

References

[1] Z. Pawlak, "Rough sets," *International Journal of Computer & Information Sciences*, vol. 11, no. 5, pp. 341-356, 1982.

[2] R. Jensen and Q. Shen, "New approaches to fuzzy-rough feature selection," *IEEE Transactions on fuzzy Systems*, vol. 17, no. 4, pp. 824-838, 2009.

[3] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*. Dordrecht: Kluwer Academic Publishers, 1991

[4] M. Kryszkiewicz, "Certain, generalized decision, and membership distribution reducts versus functional dependencies in incomplete systems," in *Rough Sets and Intelligent Systems Paradigms*. Heidelberg: Springer, 2007, pp. 162-174.

[5] M. Li, S. Deng, S. Feng, and J. Fan, "Fast assignment reduction in inconsistent incomplete decision systems," *Journal of Systems Engineering and Electronics*, vo. 25, no. 1, pp. 83-94, 2014.

[6] W. X. Zhang, J. S. Mi, and W. Z. Wu, "Approaches to knowledge reductions in inconsistent systems," *International Journal of Intelligent Systems*, vol. 18, no. 9, pp. 989-1000, 2003.

[7] M. Li, C. Shang, S. Feng, and J. Fan, "Quick attribute reduction in inconsistent decision tables," *Information Sciences*, vol. 254, pp. 155-180, 2014.

[8] M. Inuiguchi, Y. Yoshioka, and Y. Kusunoki, "Variable-precision dominance-based rough set approach and attribute reduction," *International Journal of Approximate Reasoning*, vol. 50, no. 8, pp. 1199-1214, 2009.

[9] J. Y. Wang and J. Zhou, "Research of reduct features in the variable precision rough set model," *Neurocomputing*, vol. 72, no. 10-12, pp. 2643-2648, 2009.

[10] R. R. Yager, J. Kacprzyk, and M. Fedrizzi, *Advances in the Dempster-Shafer Theory of Evidence*. New York, NY: John Wiley & Sons Inc., 1994.

[11] D. Ye and Z. Chen, "A new type of attribute reduction for inconsistent decision tables and its computation," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 18, no. 2, pp. 209-222, 2010.

[12] X. Jia, W. Liao, Z. Tang, and L. Shang, "Minimum cost attribute reduction in decision-theoretic rough set models," *Information Sciences*, vol. 219, pp. 151-167, 2013.

[13] J. Li, H. Zhao, and W. Zhu, "Fast randomized algorithm with restart strategy for minimal test cost feature selection," *International Journal of Machine Learning and Cybernetics*, vol. 6, no. 3, pp. 435-442, 2015.

[14] Q. Hu, Z. Xie, and D. Yu, "Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation," *Pattern Recognition*, vol. 40, no. 12, pp. 3509-3521, 2007.

[15] D. Q. Miao and G. R. Hu, "A heuristic algorithm for reduction of knowledge," *Journal of Computer Research and Development*, vol. 36, no. 6, pp. 681-684, 1999.

[16] G. Y. Wang, H. Yu, and D. C. Yang, "Decision table reduction based on conditional information entropy," *Chinese Journal of Computers*, vol. 25, no. 7, pp. 759-766, 2002.

[17] Z. H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL: Taylor & Francis, 2012.

[18] L. L. Minku, A. P. White, and X. Yao, "The impact of diversity on online ensemble learning in the presence of concept drift," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 5, pp. 730-742, 2010.

[19] R. Jensen and Q. Shen, "Fuzzy–rough attribute reduction with application to web categorization," *Fuzzy Sets and Systems*, vol. 141, no. 3, pp. 469-485, 2004.

[20] Y. Qian, J. Liang, W. Pedrycz, and C. Dang, "Positive approximation: an accelerator for attribute reduction in rough set theory," *Artificial Intelligence*, vol. 174, no. 9-10, pp. 597-618, 2010.

[21] M. Li, S. Deng, L. Wang, S. Feng, and J. Fan, "Hierarchical clustering algorithm for categorical data using a probabilistic rough set model," *Knowledge-Based Systems*, vol. 65, pp. 60-71, 2014.

[22] J. Ross Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 2014.

[23] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press, 2000.

[24] V. Garcia, J. S. Sanchez, and R. A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," *Knowledge-Based Systems*, vol. 25, no. 1, pp. 13-21, 2012.

Min Li https://orcid.org/0000-0001-5428-6276

He received his Ph.D. in Computer Application Technology from the Shenzhen Institute of Advanced Technology of Chinese Academy of Sciences, China in 2013. He is now a professor at the Nanchang Institute of Technology. His current research interests include data mining, knowledge discovery in database with fuzzy and rough techniques.

Shaobo Deng https://orcid.org/0000-0002-0848-5343
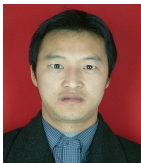
He received his Ph.D. in Computer Application Technology from Institute of Computing Technology of Chinese Academy of Sciences, China in 2015. He is now an associate professor at the Nanchang Institute of Technology. His current research interests include fuzzy and rough set theory, mathematical logic, uncertainty reasoning.

Lei Wang https://orcid.org/0000-0001-5563-8499

He received his Ph.D. in Computer Application Technology from Southwest Jiaotong University of China in 2012. He is now an associate professor at the Nanchang Institute of Technology. His current research interests include data mining, knowledge discovery in database with fuzzy and rough technique.