IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received January 10, 2019, accepted February 20, 2019, date of publication March 11, 2019, date of current version April 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2903845

Feature Selection for Partially Labeled Data Based on Neighborhood Granulation Measures

BINGYANG LI, JIANMEI XIAO, AND XIHUAI WANG
Department of Electrical Engineering, Shanghai Maritime University, Shanghai 201306, China

Corresponding author: Xihuai Wang (wxh@shmtu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61503241.

ABSTRACT As an effective feature selection technique, rough set theory plays an important part in machine learning. However, it is only applicable to labeled data. In reality, there are massive partially labeled data in machine learning tasks, such as webpage classification, speech recognition, and text categorization. To effectively remove redundant features of partially labeled data, the neighborhood granulation measures based on a neighborhood rough set model are put forward in this paper, which can be used to evaluate the discernibility ability of feature subsets under both information systems and decision systems. Moreover, a new definition of significance is introduced. Based on that, a semisupervised reduction algorithm is presented for the feature selection of partially labeled data. Several datasets are chosen to verify its effectiveness. The comparative experiments show that our proposed method is more effective and applicable to the feature selection of partially labeled data.

INDEX TERMS Partially labeled data, feature selection, neighborhood rough set, granulation measures, semisupervised learning.

I. INTRODUCTION

In recent years, since data gradually tends to be large-scale and high-dimensional, feature selection, also called attribute reduction, has attracted more and more attention. As an important part of machine learning and data mining, feature selection is a fundamental preprocessing method to eliminate redundant features or attributes. The advantages of feature selection are obvious. For example, it can effectively decrease computation burden and reduce data storage space. Moreover, the performance of learning algorithms can be improved by removing irrelevant features. Up to now, many feature selection techniques have been proposed and applied to practical applications [1]–[10], such as webpage classification [9], fault diagnosis [10], and power system transient stability assessment [6], etc.

As an important soft computing technique dealing with uncertain and vague information, one of representative characteristics of rough set model [11] is its powerful ability to handle feature selection problems. Compared with other techniques, rough set model is totally data-driven and doesn't need any other prior information [12]. However, the applications of rough set model are limited since it can only handle categorical data. Neighborhood rough set model, introduced by Yao [13], provides a feasible way to solve this problem by replacing the equivalence relation with neighborhood relation, which can avoid data discretization and directly handle real data. Many feature selection methods based on neighborhood rough set model have been proposed recently [14]–[17]. However, the current researches are mainly concentrated on labeled data and relatively less attention has been paid to the case of partially labeled data. Actually, in machine learning tasks, it's fairly common to find that only one part of sample data are labeled and the others are unlabeled, such as webpage classification, speech recognition, and image annotation, etc. [18]–[20]. This is because the labeled data are quite hard to obtain sometimes, which will cost much time or human effort. By contrast, it's much easier to acquire unlabeled data. For instance, in webpage classification, it's easy to gather the unlabeled data by a web crawler. However, it's quite expensive to obtain their decision labels since one should label these webpages manually. The existing rough set methods may be not well applicable to handle this situation due to the insufficient of labeled data. For this reason, it's necessary to research on how to take full advantage of both labeled and unlabeled data.

So far, there have been some pioneering work on the study of feature selection for partially labeled data. In [21],

The associate editor coordinating the review of this manuscript and approving it for publication was Isaac Triguero.

Miao *et al.* transformed the partially labeled data into labeled data by providing each unlabeled data with a pseudo-class symbol and designed an attribute reduction algorithm using markov blanket, but it has the weakness of poor efficiency since each unlabeled instance marked with pseudo-class symbol is required to be discernible. Similarly, Ren *et al.* [24], developed a wrapper-based feature selection method by transforming unlabeled data into labeled data. Zhu *et al.* [25] also presented a multi-label feature selection method by recovering missing data. In [22], the discernibility pair was introduced to measure different features and two reduction algorithms were proposed. In [23], an attribute reduction method using co-training was developed. However, the above methods can only handle categorical data, which have limitations in practical applications. In virtue of the current situations of feature selection for partially labeled data, as well as the need of practical applications, in this paper, we develop a feature selection method for partially labeled data using neighborhood granulation measures, which are based on neighborhood rough set model.

The contributions of this paper are three-fold. (1) Firstly, we propose the concept of neighborhood granulation measures under the framework of neighborhood rough set model, which provide a novel viewpoint to evaluate attribute subsets in both information systems and decision systems. (2) Secondly, the novel definition of significance proposed by us considers the labeled and unlabeled data simultaneously, which integrates these two parts into one without any change to original data. (3) Finally, a semisupervised feature selection algorithm is developed to handle partially labeled data. Moreover, some comparative experiments are provided to verify the effectiveness of our proposed method.

The rest of this paper is organized as follows. Some basic knowledge of partially labeled data and neighborhood rough sets are introduced in the next section. In section III, we introduce several neighborhood granulation measures and a new definition of significance is developed to evaluate features in partially labeled data. A semisupervised reduction algorithm is also proposed in this section. Section IV provides the comparative experiments and analysis. Finally, we conclude this paper in section V.

II. PRELIMINARIES

In this section, we review some basic knowledge of partially labeled decision systems and neighborhood rough sets. More details can be found in [12], [13], [21], and [22].

A. PARTIALLY LABELED DECISION SYSTEMS

In the framework of rough set theory, the research data is typically described as an information system IS $= \langle U, A, V, f\rangle$, where $U = \{x_1, x_2, \ldots, x_m\}$ is the nonempty and finite set of instances, $A = \{a_1, a_2, \ldots, a_n\}$ is the set of features (attributes), V is the set of corresponding attribute values. $f : U\times A \rightarrow V$ is an information function, which determines the value of each instance under a certain feature. IS is also called a decision system DS $= \langle U, C \cup D, V, f\rangle$, if $A = C \cup D$, where C is the set of condition attributes and D is a decision attribute. Furthermore, DS is referred to as a partially labeled decision system PDS $=\langle X_L \cup X_U, C \cup D, V, f\rangle$, if $U = X_L \cup X_U$, where X_L is the set of labeled instances and X_U is the set of unlabeled instances.

By distinguishing the three types of research data, a learning process can be further divided into supervised learning, unsupervised learning, or semisupervised learning.

B. NEIGHBORHOOD ROUGH SET MODEL

Definition 1: Let IS $= \langle U, C\rangle$ be an information system, where $U = \{x_1, x_2, \ldots, x_m\}$ is the set of instances, $C = \{c_1, c_2, \ldots, c_n\}$ is the set of condition attributes, $B\subseteq C$. δ is a nonnegative number ranging in [0, 1]. Then, the neighborhood relation induced by B is defined as

$$NR_B^\delta = \left\{(x_i, x_j) \in U \times U | \Delta_B(x_i, x_j) \leq \delta\right\} \tag{1}$$

where δ is a neighborhood threshold which controls the neighborhood size of each instance, Δ_B is a distance function, usually denoted by Minkowski Distance as follows:

$$\Delta_B(x_i, x_j) = \left(\sum\nolimits_{\forall c\in B} \left|f\ (x_i, c) - f(x_j, c)\right|^P\right)^{1/P} \tag{2}$$

$\forall x_i, x_j, x_k \in U$, the distance function satisfies nonnegativity ($\Delta_B(x_i, x_j) \geq 0$), symmetry ($\Delta_B\left(x_i, x_j\right) = \Delta_B\left(x_j, x_i\right)$) and transitivity ($\Delta_B\left(x_i, x_k\right) \leq \Delta_B\left(x_i, x_j\right) + \Delta_B\left(x_j, x_k\right)$).

For convenience, hereinafter, we denote NR_B^δ by NR_B if there is no confusion.

Definition 2: Let IS $=\langle U, C\rangle$ be an information system, where $U = \{x_1, x_2, \ldots, x_m\}$ is the set of instances, C is the set of condition attributes, $B \subseteq C$. The neighborhood relation determined by B is NR_B^δ, then the neighborhood granulation of U induced by NR_B^δ is denoted by

$$U/NR_B^\delta = \{n_B^\delta\left(x_1\right), n_B^\delta\left(x_2\right), \ldots, n_B^\delta\left(x_m\right)\} \tag{3}$$

where $n_B^\delta\left(x_i\right) = \{x_j | (x_i, x_j) \in NR_B^\delta\}$ denotes the neighborhood granule of x_i in feature B.

It's noted that the equivalence relation in traditional rough sets constitutes a partition of U, whereas the neighborhood relation in neighborhood rough sets forms a covering of U. The neighborhood granulation of U induced by NR_B is also called a neighborhood granular structure. For convenience, in what follows, we denote $n_B^\delta\left(x_i\right)$ by $n_B(x_i)$, if no confusion arises.

Theorem 1: Let IS $=\langle U, C\rangle$ be an information system, $B_1, B_2\subseteq C$, $B_1 \subseteq B_2$. δ is a neighborhood threshold. Then for any $x_i \in U$, there exists $n_{B_2}^\delta\left(x_i\right) \subseteq n_{B_1}^\delta\left(x_i\right)$.

Theorem 2: Let IS $=\langle U, C\rangle$ be an information system, $B \subseteq C$. δ_1, δ_2 are two neighborhood thresholds satisfying $\delta_1 \leq \delta_2$. Then for any $x_i \in U$, there exists $n_B^{\delta_1}\left(x_i\right) \subseteq n_B^{\delta_2}\left(x_i\right)$.

Theorem 1 and theorem 2 show that adding new features or decreasing the value of neighborhood threshold will lead to the diminution of neighborhood granules.

Definition 3: Let DS $=\langle U, C \cup D\rangle$ be a decision system, where $U = \{x_1, x_2, \ldots, x_m\}$ is the set of instances, $C = \{c_1, c_2, \ldots, c_n\}$ is the set of condition attributes, $B\subseteq C, D$

is the decision attribute. $d_1, d_2, \ldots, d_N$ are the equivalence classes obtained by D. Then the lower and upper approximations of D with regard to B are defined as follows:

$$\underline{NR_B}D = \bigcup\nolimits_{i=1}^{N} \underline{NR_B}d_i; \overline{NR_B}D = \bigcup\nolimits_{i=1}^{N} \overline{NR_B}d_i$$

where $\underline{NR_B}d_i = \left\{x_j | n_B\left(x_j\right) \subseteq d_i, x_j \in U\right\}$, and $\overline{NR_B}d_i = \{x_j | n_B\left(x_j\right) \cap d_i \neq \emptyset, x_j \in U\}$.

The lower approximation of D with respect to B is also called the positive region of D with respect to B, denoted by $POS_B(D)$.

Definition 4: Let DS $=\langle U, C \cup D\rangle$ be a decision system, $B \subseteq C$, then B is referred to as a reduct if B satisfies $POS_B\left(D\right) = POS_C(D)$, and $\forall B' \subset B$, there exists $POS_{B'}\left(D\right) < POS_B\left(D\right)$.

III. FEATURE SELECTION FOR PARTIALLY LABELED DATA BASED ON NEIGHBORHOOD GRANULATION MEASURES

From the viewpoint of neighborhood rough sets, information or knowledge is implied in attributes, which generate a series of neighborhood granules of referential universe and constitute a covering of the referential universe. Generally speaking, adding new features will make the neighborhood granules finer and the objects can be approximated more accurately. Namely, the finer the granule is, the stronger the discernibility power will be. Thus, knowledge can be measured by studying the information granular structures induced by corresponding attribute subsets. As one of important uncertainty measures, granulation measures have been thoroughly researched and applied to many applications in recent years [26]–[29]. In particular, Liang and Qian [26] introduced the information granulation for information systems. In [28], the combination information granulation was researched. In this section, we will extend the granulation measures into neighborhood granular structure. Based on that, a novel concept of significance will be developed to measure attributes in partially labeled decision system.

A. NEIGHBORHOOD GRANULATION MEASURES

Definition 5: Let IS $=\langle U, C\rangle$ be an information system, $B \subseteq C. U/NR_B^{\delta} = \{n_B^{\delta}\left(x_1\right), n_B^{\delta}\left(x_2\right), \ldots, n_B^{\delta}\left(x_{|U|}\right)\}$ is a neighborhood granulation of U induced by B. δ is a neighborhood threshold. Then the neighborhood information granulation of B is defined as follows:

$$NG^{\delta}\left(B\right) = \frac{1}{|U|^2} \sum\nolimits_{i=1}^{|U|} |n_B^{\delta}(x_i)| \tag{4}$$

If for any $x_i \in U$, there exists $\left|n_B^{\delta}\left(x_i\right)\right| = |U|$, then $NG^{\delta}\left(B\right) = 1$. If for any $x_i \in U$, there exists $\left|n_B^{\delta}\left(x_i\right)\right| = 1$, then $NG^{\delta}\left(B\right) = 1/|U|$. Since $NG^{\delta}\left(B\right)$ increases monotonically with $|n_B^{\delta}(x_i)|$, we can obtain that $1/|U| \leq NG^{\delta}\left(B\right) \leq 1$. For convenience, in what follows, we denote $NG^{\delta}\left(B\right)$ by $NG\left(B\right)$, if there is no confusion.

Theorem 3: Let IS $=\langle U, C\rangle$ be an information system, $B_1, B_2 \subseteq C, B_1 \subseteq B_2$, then $NG\left(B_2\right) \leq NG\left(B_1\right)$.

Proof: By theorem 1, we can obtain that $n_{B_2}\left(x_i\right) \subseteq n_{B_1}\left(x_i\right)$, which means that $|n_{B_2}(x_i)| \leq |n_{B_1}(x_i)|$. Hence, $NG\left(B_2\right) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} |n_{B_2}(x_i)| \leq NG\left(B_1\right) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} |n_{B_1}(x_i)|$.

Theorem 4: Let IS $=\langle U, C\rangle$ be an information system, $B \subseteq C$. δ_1, δ_2 are two neighborhood thresholds, $\delta_1 \leq \delta_2$, then $NG^{\delta_1}\left(B\right) \leq NG^{\delta_2}\left(B\right)$.

Proof: By theorem 2, we can obtain that $n_B^{\delta_1}\left(x_i\right) \subseteq n_B^{\delta_2}\left(x_i\right)$, which means that $|n_B^{\delta_1}\left(x_i\right)| \leq |n_B^{\delta_2}\left(x_i\right)|$. Hence, $NG^{\delta_1}\left(B\right) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} |n_B^{\delta_1}(x_i)| \leq NG^{\delta_2}\left(B\right) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} |n_B^{\delta_2}(x_i)|$.

Theorem 3 shows that for a given feature set, the neighborhood information granulation decreases monotonously as new features are added into the feature set. Thus, the neighborhood information granulation can be used to measure the discernibility ability of different granular structures in an information system. Theorem 4 shows that the neighborhood threshold also has an impact on the neighborhood information granulation. Based on definition 5 and theorem 3, the concept of attribute reduct for an information system can be developed.

Definition 6: Let IS $=\langle U, C\rangle$ be an information system, $B \subseteq C$. Then B is called a reduct of IS, if there exist:

(1) $NG\left(B\right) = NG\left(C\right)$;

(2) $\forall B' \subset B, NG\left(B'\right) > NG\left(B\right)$.

In the above definition, condition (1) guarantees that the reduct owns identical discernibility power with the original feature set. While condition (2) ensures that there is no redundancy in the chosen reduct set.

The neighborhood information granulation provides us an index for attribute evaluation. Further, we can use forward or backward feature selection strategy to find the reduct set. In the forward feature selection strategy, we start with an empty set. In each round, the feature which maximizes the decrement of neighborhood information granulation of the current set will be added into reduct set and the algorithm will keep running until the neighborhood information granulation of the current set is equal to the one of full set. As to the backward feature selection strategy, we first start with the full feature set. In each round, we compute the significance of each feature c_i in the current set RED by $NG\left(RED - c_i\right) - NG\left(RED\right)$, and the feature with minimum significance will be removed from the current set. At last, the algorithm will stop if the value of minimum significance is greater than zero. Both of forward and backward feature selection strategies can be utilized to find the reduct set. In this paper, the forward feature selection is adopted. Then, we develop a feature selection algorithm for information systems (unlabeled data) based on neighborhood information granulation as shown in algorithm 1.

It's noted that neighborhood information granulation is based on the assumption of single neighborhood relation. To measure the neighborhood granular structure induced by multiple neighborhood relations, the following concepts are introduced.

Definition 7: Let DS $=\langle U, C \cup D\rangle$ be a decision system, $B_1, B_2 \subseteq C. U/NR_{B_1} = \{n_{B_1}\left(x_1\right), n_{B_1}\left(x_2\right), \ldots, n_{B_1}\left(x_{|U|}\right)$

Algorithm 1 Attribute Reduction for Information Systems Based on Neighborhood Granulation Measures (NGU for Short)

```
Input: an information system IS $=\langle U, C\rangle$, $RED=\emptyset$.
Output: the reduct set $RED$.
Step 1: normalize the data by min-max normalization.
Step 2: for each $a_i \in C - RED$,
            compute $NG\left(RED \cup \{a_i\}\right)$.
        end for
Step 3: find $a_l$ satisfying
        $NG\left(RED \cup \{a_l\}\right) = min_i NG\left(RED \cup \{a_i\}\right)$.
Step 4: $RED = RED \cup a_l$.
Step 5: if $NG\left(RED\right) \neq NG\left(C\right)$
          go to step 2.
        else
          go to step 6.
        end if
Step 6: return the reduct set $RED$.
```

and $U/NR_{B_2} = \{n_{B_2}(x_1), n_{B_2}(x_2), \ldots, n_{B_2}(x_{|U|})\}$ are neighborhood granulations of U induced by B_1 and B_2 respectively. Then the neighborhood combination information granulation of B_1 and B_2 is defined as follows:

$$NG\left(B_1, B_2\right) = \frac{1}{|U|^2} \sum\nolimits_{i=1}^{|U|} |n_{B_1}(x_i) \cap n_{B_2}(x_i)| \tag{5}$$

Furthermore, let $U/R_D = \{[x_1]_D, [x_2]_D, \ldots, [x_{|U|}]_D\}$, where R_D is the equivalence relation induced by decision attribute D and $[x_i]_D$ is the equivalence class of x_i generated by R_D. Then, $\forall B \subseteq C$, the neighborhood combination information granulation of D and B is defined as follows:

$$NG\left(D, B\right) = \frac{1}{|U|^2} \sum\nolimits_{i=1}^{|U|} |[x_i]_D \cap n_B(x_i)| \tag{6}$$

Theorem 5: Let IS $=\langle U, C\rangle$ be an information system, $B_1, B_2 \subseteq C$, then there exist $NG\left(B_1, B_2\right) \leq NG\left(B_1\right)$ and $NG\left(B_1, B_2\right) \leq NG\left(B_2\right)$.

Proof: Since $\left(n_{B_1}\left(x_i\right) \cap n_{B_2}\left(x_i\right)\right) \subseteq n_{B_1}(x_i)$, it's easy to obtain that $|n_{B_1}(x_i) \cap n_{B_2}(x_i)| \leq |n_{B_1}(x_i)|$. Therefore, we can conclude that $NG\left(B_1, B_2\right) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} |n_{B_1}(x_i) \cap n_{B_2}(x_i)| \leq NG\left(B_1\right) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} |n_{B_1}(x_i)|$. It's similar to prove that $NG\left(B_1, B_2\right) \leq NG\left(B_2\right)$.

Theorem 6: Let IS $=\langle U, C\rangle$ be an information system, $B \subseteq C$. δ_1, δ_2 are two neighborhood thresholds, $\delta_1 \leq \delta_2$, then $NG^{\delta_1}\left(B_1, B_2\right) \leq NG^{\delta_2}\left(B_1, B_2\right)$.

Proof: By theorem 2, we can obtain that $\left(n_{B_1}^{\delta_1}\left(x_i\right) \cap n_{B_2}^{\delta_1}\left(x_i\right)\right) \subseteq \left(n_{B_1}^{\delta_2}\left(x_i\right) \cap n_{B_2}^{\delta_2}\left(x_i\right)\right)$, which means that $|n_{B_1}^{\delta_1}\left(x_i\right) \cap n_{B_2}^{\delta_1}\left(x_i\right)| \leq |n_{B_1}^{\delta_2}\left(x_i\right) \cap n_{B_2}^{\delta_2}\left(x_i\right)|$. Hence, $NG^{\delta_1}\left(B_1, B_2\right) \leq NG^{\delta_2}\left(B_1, B_2\right)$.

The above theorems show that the neighborhood combination information granulation of two attribute subsets is less than the single one. Moreover, the smaller the neighborhood threshold is, the smaller the neighborhood combination information granulation will be.

Definition 8: Let DS $=\langle U, C \cup D\rangle$ be a decision system, $B_1, B_2 \subseteq C. U/NR_{B_1} = \{n_{B_1}(x_1), n_{B_1}(x_2), \ldots, n_{B_1}(x_{|U|})\}$ and $U/NR_{B_2} = \{n_{B_2}(x_1), n_{B_2}(x_2), \ldots, n_{B_2}(x_{|U|})\}$ are neighborhood granulations of U induced by B_1 and B_2 respectively. Then the conditional neighborhood information granulation of B_2 with respect to B_1 is defined as follows:

$$NG\left(B_2 \,|\, B_1\right) = \frac{1}{|U|^2} \sum\nolimits_{i=1}^{|U|} \left(\left|n_{B_1}\left(x_i\right)\right| - \left|n_{B_1}\left(x_i\right) \cap n_{B_2}\left(x_i\right)\right|\right) \tag{7}$$

Furthermore, let $U/R_D = \{[x_1]_D, [x_2]_D, \ldots, [x_{|U|}]_D\}$. Then, $\forall B \subseteq C$, the conditional neighborhood information granulation of D with respect to B is defined as follows:

$$NG\left(D \,|\, B\right) = \frac{1}{|U|^2} \sum\nolimits_{i=1}^{|U|} \left(|n_B\left(x_i\right)| - |n_B\left(x_i\right) \cap [x_i]_D|\right) \tag{8}$$

Theorem 7: Let DS $=\langle U, C \cup D\rangle$ be a decision system, $B_1, B_2 \subseteq C$. Then $NG\left(B_2 \,|\, B_1\right) = NG\left(B_1\right) - NG\left(B_1, B_2\right)$.

Proof: It follows directly from definitions 5-7.

Theorem 8: Let DS $=\langle U, C \cup D\rangle$ be a decision system, $B_1, B_2 \subseteq C, B_1 \subseteq B_2$. Then $NG\left(B_1 \,|\, B_2\right) = 0$.

Proof: By theorem 1, we can obtain that $n_{B_2}\left(x_i\right) \subseteq n_{B_1}\left(x_i\right)$, which means that $\left|n_{B_2}\left(x_i\right)\right| = |n_{B_1}(x_i) \cap n_{B_2}\left(x_i\right)|$. Hence, $NG\left(B_1 \,|\, B_2\right) = \frac{1}{|U|^2} \sum_{i=1}^{|U|}\left(\left|n_{B_2}\left(x_i\right)\right| - |n_{B_1}\left(x_i\right) \cap n_{B_2}\left(x_i\right)|\right) = 0$.

Theorem 9: Let DS $=\langle U, C \cup D\rangle$ be a decision system, $B_1, B_2 \subseteq C, B_1 \subseteq B_2$. Then $NG\left(D \,|\, B_2\right) \leq NG\left(D \,|\, B_1\right)$.

Proof: By theorem 1, we can obtain that $n_{B_2}\left(x_i\right) \subseteq n_{B_1}\left(x_i\right)$, which means that $|n_{B_2}(x_i)| \leq |n_{B_1}(x_i)|$. Hence, $NG\left(D \,|\, B_2\right) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} \left(\left|n_{B_2}\left(x_i\right)\right| - |n_{B_2}\left(x_i\right) \cap [x_i]_D|\right) \leq NG\left(D \,|\, B_1\right) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} \left(\left|n_{B_1}\left(x_i\right)\right| - |n_{B_1}\left(x_i\right) \cap [x_i]_D|\right)$.

Theorem 7 shows that the conditional neighborhood information granulation is the information increment between neighborhood information granulation and combination information granulation, which reflects correlations between different attribute subsets. Theorem 8 shows that for a given attribute set, the information increment provided by its arbitrary subset is zero. Theorem 9 shows that adding new condition attributes will not increase the conditional information granulation of decision attribute with respect to the condition attribute subset. Namely, the more attributes there are, the lower the conditional information granulation is and the stronger the discernibility ability of an attribute subset will be. Thus, the concept of attribute reduct for a decision system (labeled data) can be defined as follows.

Definition 9: Let DS $=\langle U, C \cup D\rangle$ be a decision system, $B \subseteq C$. Then B is called a reduct of DS, if there exist:

(1) $NG\left(D \,|\, B\right) = NG\left(D \,|\, C\right)$;

(2) $\forall B' \subset B, NG\left(D \,|\, B'\right) > NG\left(D \,|\, B\right)$.

Further, we develop a feature selection algorithm for decision system (labeled data) based on conditional neighborhood information granulation as shown in algorithm 2.

Algorithm 2 Attribute Reduction for Decision Systems Based on Neighborhood Granulation Measures (NGL for Short).

```
Input: a decision system DS $=\langle U, C \cup D\rangle$, $RED = \emptyset$.
Output: the reduct set $RED$.
Step 1: normalize the data by min-max normalization.
Step 2: for each $a_i \in C - RED$
            compute $NG\,(D|RED \cup \{a_i\})$.
        end for
Step 3: find the attribute $a_l$ satisfying
        $NG\,(D|RED \cup \{a_l\}) = min_i NG\,(D|RED \cup \{a_i\})$.
Step 4: $RED = RED \cup a_l$.
Step 5: if $NG\,(D|RED) \neq NG\,(D|C)$
            turn to Step 2
        else
            turn to Step 6
        end if
Step 6: return the reduct set $RED$.
```

B. ATTRIBUTE REDUCTION FOR PARTIALLY LABELED DATA

In the previous subsection, we introduce two feature selection algorithms for labeled data (decision systems) and unlabeled data (information systems) respectively, which are actually under the framework of supervised learning and unsupervised learning respectively. In practice, there usually exist massive partially labeled data in semisupervised learning tasks. However, most current studies of rough set theory about feature selection techniques mainly concentrate on labeled data and there is little discussion with regard to partially labeled data. Therefore, in this subsection, we propose a feature selection method to handle partially labeled data.

At present, there are usually two strategies for handling partially labeled data. One is transforming partially labeled data into labeled data. However, this method has the weakness of poor efficiency. Moreover, the process of transformation may increase the computation complexity. The other one directly uses the supervised or unsupervised methods to handle partially labeled data. Whereas this strategy may lead to poor performance due to the lack of effective learning samples. To ensure the learning performance, we should fully use both unlabeled and labeled data. From the previous discussion, we know that the neighborhood granulation measures could deal with both supervised and unsupervised data. Thus, we could use the neighborhood granulation measures to evaluate the labeled part and unlabeled part simultaneously. Based on the above analysis, we develop the following concept.

Definition 10: Let PDS $= \langle X_L \cup X_U, C \cup D\rangle$ be a partially labeled decision system, where X_L is the set of labeled instances and X_U is the set of unlabeled instances. $\forall B \subseteq C$, the significance of condition attribute set B with respect to PDS is defined as follows:

$$SIG\,(B) = \begin{cases} \dfrac{NG_U\,(C)}{NG_U\,(B)} + \dfrac{1+NG_L(D|C)}{1+NG_L(D|B)}, & if\ X_L \neq \emptyset, X_U \neq \emptyset \\ \dfrac{NG_U\,(C)}{NG_U\,(B)}, & if\ X_L = \emptyset, X_U \neq \emptyset \\ \dfrac{1+NG_L(D|C)}{1+NG_L(D|B)}, & if\ X_L \neq \emptyset, X_U = \emptyset \end{cases} \tag{9}$$

where $NG_U\,(B)$ denotes the neighborhood information granulation of B computed by the part of unlabeled data X_U and $NG_L(D|B)$ denotes the conditional neighborhood information granulation of D with respect to B computed by the part of labeled data X_L.

In definition 10, the unlabeled part is measured by the neighborhood information granulation and the labeled part is measured by the conditional neighborhood information granulation. The significance measure integrates these two parts into one. Actually, it's quite clear to use the concept of significance. If one feature subset has the same significance with the original feature set, then it has the same discernibility power since the neighborhood granular structure is not changed. Hence, we can use the proposed significance to design attribute reduct for partially labeled data.

Definition 11: Let PDS $= \langle X_L \cup X_U, C \cup D\rangle$ be a partially labeled decision system, where X_L is the set of labeled instances and X_U is the set of unlabeled instances. $B \subseteq C$. Then B is called a reduct of PDS, if there exist:

(1) $SIG\,(B) = SIG\,(C)$;
(2) $\forall B' \subset B, SIG\left(B'\right) < SIG\,(B)$.

As mentioned above, the significance reflects the discernibility power of one condition attribute subset. The greater the significance is, the stronger the discernibility power of the attribute subset is. Thus, we could start with an empty reduct set and pick up the attribute with maximum increment of significance of the present reduct set in each round. Based on the above analysis, we further construct a feature selection algorithm for partially labeled decision system (partially labeled data) based on the proposed significance measure which is shown in algorithm 3.

In algorithm 3, the computation complexity mainly determined by Step 2-Step 4. Suppose that there are n condition attributes, m_l labeled instances, and m_u unlabeled instances. $m_l + m_u = m$. Then the computation complexity of computing significance of attribute set is equal to $O(m_l^2 n + m_u^2 n)$. In a worst case, the computation complexity of updating reduct set (i.e. Step 2-Step 5) is $O(m_l^2 n^2 + m_u^2 n^2)$. Thus, the overall computation complexity of algorithm 3 is equal to $O(m_l^2 n^2 + m_u^2 n^2)$.

Algorithm 3 Attribute Reduction for Partially Labeled Decision Systems Based on Neighborhood Granulation Measures (NGAR for Short).

```
Input: partially labeled decision system PDS $=\langle X_L \cup X_U, C \cup D\rangle$, $RED = \emptyset$.
Output: reduct set $RED$.
Step 1: normalize the data by min-max normalization.
Step 2: for each $a_i \in C - RED$
            compute $SIG\,(RED \cup \{a_i\})$.
        end for
Step 3: find the attribute $a_l$ satisfying
        $SIG\,(RED \cup \{a_l\}) = max_i SIG\,(RED \cup \{a_i\})$.
Step 4: $RED = RED \cup a_l$.
Step 5: If $SIG\,(RED) \neq SIG\,(C)$,
            turn to Step 2.
        else
            turn to Step 6.
        end if
Step 6: Return the reduct set $RED$.
```

TABLE 1. A partially labeled decision system.

U	c_1	c_2	c_3	c_4	c_5	D
x_1	0	0.73	0.25	0.93	0.18	1
x_2	0	0.77	0.08	0.96	0.15	1
x_3	0.28	0.73	1	0.95	0.16	*
x_4	0.54	0.85	0	0.72	0	*
x_5	0.77	0.85	0.50	1	0.12	2
x_6	0.66	0.88	0.83	0.99	0.22	2
x_7	0.94	1	0.25	0.48	0.47	*
x_8	1	0	0.50	0	1	3

C. AN ILLUSTRATIVE EXAMPLE

In this subsection, an illustrative example is employed here to show the rationale and main steps of our proposed feature selection method for partially labeled data.

Consider a partially labeled decision system PDS $=\langle X_L \cup X_U, C \cup D\rangle$ shown in Table 1, where $X_L = \{x_1, x_2, x_5, x_6, x_8\}$, $X_U = \{x_3, x_4, x_7\}$, $C = c_1, c_2, c_3, c_4, c_5\}$. Assume $\delta = 0.35$. The Euclidean distance is adopted here.

1) First, by NGAR, we have $n_C\,(x_1) = \{x_1, x_2\}$, $n_C\,(x_2) = \{x_1, x_2\}$, $n_C\,(x_3) = \{x_3\}$, $n_C\,(x_4) = \{x_4\}$, $n_C\,(x_5) = \{x_5\}$. Thus, by definition 5 and 8, we can obtain that $NG_L\,(D|C) = 0$, $NG_U\,(C) = 0.33$. In the first loop, $SIG\,(c_1) = \frac{0.33}{0.56} + \frac{1+0}{1+0.16} = 1.46$, $SIG\,(c_2) = \frac{0.33}{1} + \frac{1+0}{1+0.32} = 1.09$, $SIG\,(c_3) = \frac{0.33}{0.56} + \frac{1+0}{1+0.32} = 1.36$, $SIG\,(c_4) = \frac{0.33}{0.78} + \frac{1+0}{1+0.32} = 1.19$, $SIG\,(c_5) = \frac{0.33}{0.78} + \frac{1+0}{1+0.32} = 1.19$. Thus, c_1 is put into reduct set, namely, $RED = \{c_1\}$. In the second loop, $SIG\,(RED \cup \{c_2\}) = \frac{0.33}{0.56} + \frac{1+0}{1+0} = 1.6$, $SIG\,(RED \cup \{c_3\}) = \frac{0.33}{0.33} + \frac{1+0}{1+0.08} = 1.93$, $SIG\,(RED \cup \{c_4\}) = \frac{0.33}{0.56} + \frac{1+0}{1+0} = 1.6$, $SIG\,(RED \cup \{c_5\}) = \frac{0.33}{0.56} + \frac{1+0}{1+0} = 1.6$. Thus, c_3 is put into reduct set, namely, $RED = \{c_1, c_3\}$. In the third loop, $SIG\,(RED \cup \{c_2\}) = \frac{0.33}{0.33} + \frac{1+0}{1+0} = 2$, $SIG\,(RED \cup \{c_4\}) = \frac{0.33}{0.33} + \frac{1+0}{1+0} = 2$, $SIG\,(RED \cup \{c_5\}) = \frac{0.33}{0.33} + \frac{1+0}{1+0} = 2$. Thus, c_2 is put into reduct set and the final reduct set is $RED = \{c_1, c_3, c_2\}$.

2) On the other hand, if we use NGU to find reduct set, then in the first loop, $NG_U\,(c_1) = 0.56$, $NG_U\,(c_2) = 1$, $NG_U\,(c_3) = 0.56$, $NG_U\,(c_4) = 0.78$, $NG_U\,(c_5) = 0.78$. Thus, c_1 is put into reduct set, namely, $RED = \{c_1\}$. In the second loop, $NG_U\,(RED \cup \{c_2\}) = 0.56$, $NG_U\,(RED \cup \{c_3\}) = 0.33$, $NG_U\,(RED \cup \{c_4\}) = 0.56$, $NG_U\,(RED \cup \{c_5\}) = 0.56$. Since $NG_U\,(RED \cup \{c_3\}) = NG_U\,(C)$, c_3 is put into reduct set and the final reduct set is $RED = \{c_1, c_3\}$.

3) If we further use NGL to find reduct set, then in the first loop, $NG_L\,(D\,|\,c_1) = 0.16$, $NG_L\,(D\,|\,c_2) = 0.32$, $NG_L\,(D\,|\,c_3) = 0.32$, $NG_L\,(D\,|\,c_4) = 0.32$, $NG_L\,(D\,|\,c_5) = 0.32$. Thus, c_1 is put into reduct set, namely, $RED = \{c_1$. In the second loop, since $NG_L\,(D\,|\,RED \cup \{c_2\}) = 0$, $NG_L\left(D\,|\,RED \cup \{c_3\}\right) = 0.08$, $NG_L\,(D\,|\,RED \cup \{c_4\}) = 0$, $NG_L\left(D\,|\,RED \cup \{c_5\}\right) = 0$, c_2 is put into reduct set and the final reduct set is $RED = \{c_1, c_2\}$.

It can be seen that the results obtained by these three methods are quite different. The underlying reason is that NGU and NGL only consider particular part of instances, which may sacrifice the learning performance due to the lack of sample knowledge. Whereas NGAR takes full advantage of both labeled and unlabeled instances, which may be more effective to deal with partially labeled data.

IV. EXPERIMENTAL ANALYSIS

In this section, to verify the effectiveness of our proposed method, a series of experiments are conducted on several datasets chosen from UCI Machine Learning Repository [30], which are shown in Table 2. There are totally three parts considered in this section. In part I, we analyze the performance of our proposed method by comparing with several supervised and unsupervised methods. In the second part, the classification performance and reduction rate of our proposed feature selection method are compared with several existing feature selection methods designed for partially labeled data. Finally, in part III, we discuss the impact of neighborhood threshold on the performance of our proposed method.

TABLE 2. Data description.

ID	Data	Features	Class	Instances
1	Wine	13	3	178
2	Iono	34	2	351
3	Glass	10	7	214
4	Wpbc	33	2	198
5	Wdbc	31	2	569
6	ILPD	10	2	583
7	Sonar	60	2	208
8	Climate	18	2	540

In the following experiments, all the numerical attributes are normalized by min-max normalization to eliminate the influence caused by the difference of units of measures. Three typical classifiers including classification and regression tree

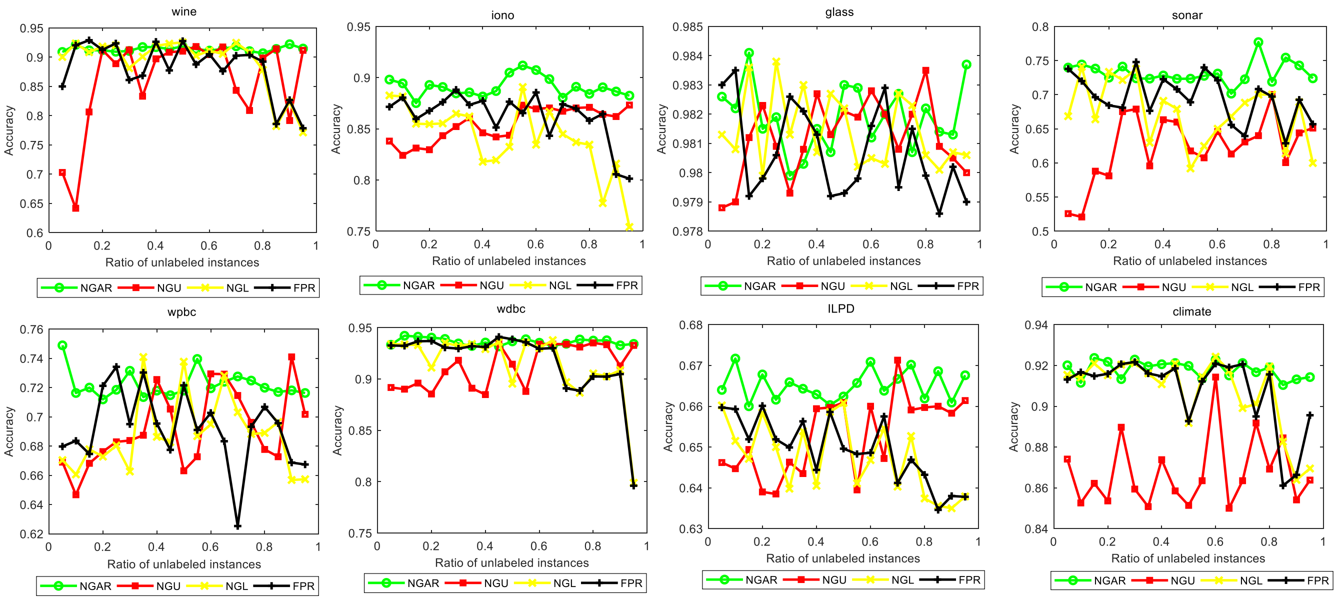

FIGURE 1. Classification performance with different label rates (CART).

(CART), radius basis function-based support vector machine (RBF-SVM) and Naïve Bayes (NB) are utilized to evaluate the classification performance of corresponding reduct sets with 10-fold cross validation scheme. For each dataset, we generate the partially labeled data by randomly selecting part of instances and removing their decision labels. The above experiments are all carried out in MATLAB 2013b.

A. COMPARISONS WITH SUPERVISED AND UNSUPERVISED FEATURE SELECTION METHODS

As mentioned above, when handling partially labeled data, one widely used strategy is to directly use supervised methods or unsupervised methods. To illustrate the necessity of our work, in this subsection, we first compare our method with several supervised and unsupervised methods, including NGU, NGL and the forward feature selection method based on positive region [31] (FPR for short).

In the above-mentioned methods, NGU is unsupervised method, which is applicable to information systems (unlabeled data). Whereas NGL and FPR are supervised methods, which are designed for decision systems (labeled data). In NGL, the conditional neighborhood information granulation is used to measure the discernibility power of different attribute sets. As to FPR, the positive region is utilized to evaluate the discernibility power of one attribute set and the attribute which can maximum the increment of positive region will be added into the reduct set until it has the same positive region with the original attribute set.

In order to test the performance of our proposed method on partially labeled data, we conduct experiments on each dataset with different ratios of unlabeled instances, where the ratio of unlabeled instances is set as 0.05 to 0.95 with the step of 0.05. Since different classifiers learn data from distinct ways, they usually need different attribute subsets to achieve the optimal classification performance. Thus, we change the value of neighborhood threshold and choose the reduct set with optimal classification performance in CART, RBF-SVM and NB respectively as the final reduct set. Figs 1-3. present the classification performance of four algorithms with different ratios of unlabeled instances. The average classification accuracies of these algorithms in different ratios of unlabeled instances are reported in Tables 3-5.

TABLE 3. Average classification performance (CART).

Data	NGAR	NGU	NGL	FPR
Wine	**0.9140**	0.8591	0.8910	0.8819
Iono	**0.8910**	0.8543	0.8411	0.8626
Glass	**0.9819**	0.9812	0.9814	0.9807
Wpbc	**0.7220**	0.6918	0.6880	0.6919
Wdbc	**0.9362**	0.9131	0.9145	0.9169
ILPD	**0.6651**	0.6528	0.6475	0.6494
Sonar	**0.7322**	0.6231	0.6733	0.6949
Climate	**0.9179**	0.8685	0.9073	0.9080
average	**0.8450**	0.8055	0.8180	0.8233

TABLE 4. Average classification performance (RBF-SVM).

Data	NGAR	NGU	NGL	FPR
Wine	**0.9509**	0.8921	0.9083	0.9001
Iono	**0.9232**	0.8885	0.8267	0.8615
Glass	**0.9223**	0.9190	0.9117	0.9143
Wpbc	**0.7175**	0.6808	0.6466	0.6367
Wdbc	**0.9625**	0.9371	0.9420	0.9477
ILPD	**0.6445**	0.6215	0.6145	0.6136
Sonar	**0.7849**	0.6926	0.7083	0.7557
Climate	**0.9271**	0.8280	0.8912	0.8630
average	**0.8541**	0.8075	0.8062	0.8116

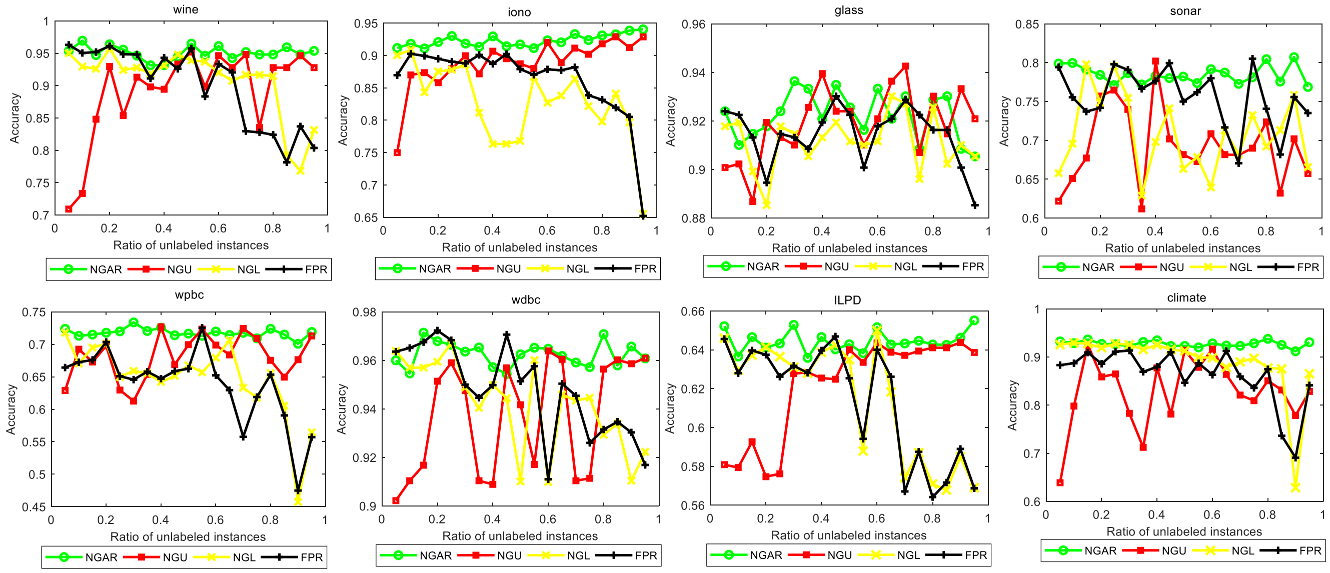

FIGURE 2. Classification performance with different label rates (RBF-SVM).

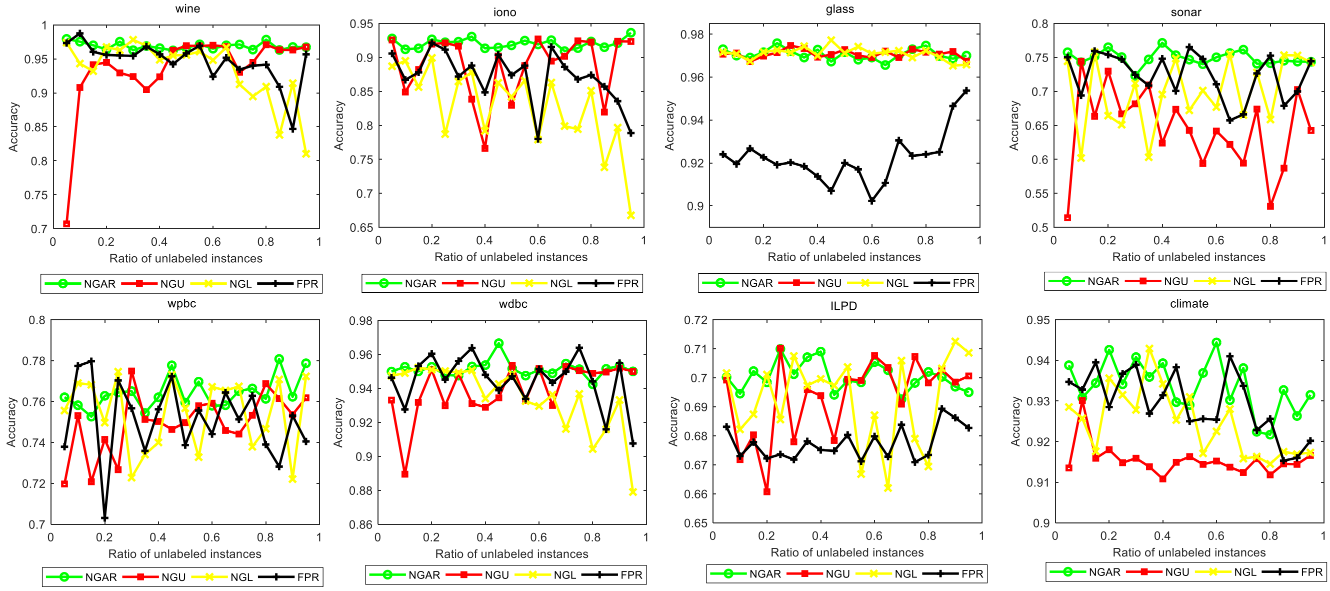

FIGURE 3. Classification performance with different label rates (Naïve Bayes).

The results in Figs 1-3. show that our proposed method has better classification performance. Moreover, compared with other methods, the classification performance of NGAR is more stable, which means that our reduction method is more reliable. As to NGU, we can find that the classification performance is poor as a relatively low percentage of unlabeled data is adopted. Although the classification performance tends to increase with the percentage of unlabeled data, the classification performance of NGU is quite unstable in general. The underlying reason is that it fails to associate condition attributes with the decision information. As to NGL and FPR, we can observe that the classification accuracies decrease with the grow of ratio of unlabeled data, especially when it is greater than 0.6. As we mentioned above, since NGL and FPR can only measure labeled data, the unlabeled data would be discarded, which will lead to the waste of unlabeled data and the lack of training data. Thus, they are not well applicable to handling partially labeled data. In contrast, our proposed method considers both labeled data and unlabeled data. On the one hand, it can take full advantage of decision information, which avoids the drawback of unsupervised methods. On the other hand, the unlabeled information can provide more useful knowledge, which avoids the problem of sacrifice of training data in supervised methods. From the

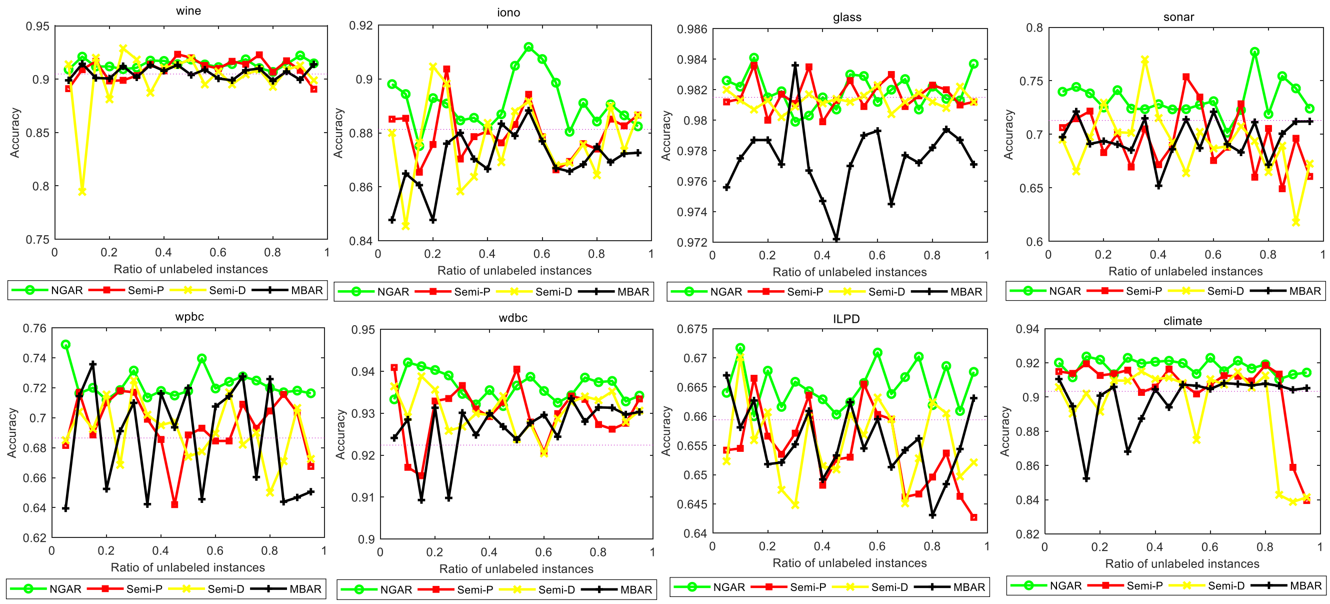

FIGURE 4. Comparisons of classification performance with existing reduction methods for partially labeled data (CART).

TABLE 5. Average classification performance (NB).

Data	NGAR	NGU	NGL	FPR
Wine	**0.9688**	0.9348	0.9337	0.9466
Iono	**0.9205**	0.8882	0.8273	0.8717
Glass	0.9706	0.9708	**0.9710**	0.9223
Wpbc	**0.7642**	0.7495	0.7540	0.7509
Wdbc	**0.9514**	0.9395	0.9347	0.9448
ILPD	**0.7004**	0.6935	0.6925	0.6774
Sonar	**0.7490**	0.6441	0.6991	0.7229
Climate	**0.9337**	0.9154	0.9245	0.9294
average	**0.8698**	0.8420	0.8421	0.8458

above analysis, it can be seen that compared with supervised and unsupervised reduction methods, our proposed method is more applicable to handling partially labeled data.

B. COMPARISONS WITH OTHER REDUCTION METHODS FOR PARTIALLY LABELED DATA

In this subsection, we compare the effectiveness of our proposed method with some existing reduction methods for partially labeled data, including semi-D [22], which are based on discernibility matrix and discernibility pairs, semi-P [22], which uses dependency degree and discernibility pairs to evaluate attributes in partially labeled data, and markov blanket-based attribute reduction method (MBAR) [21], which uses markov blanket to find optimal reduct set. Since the above three methods are only applicable to categorical values, the equal frequency discretization is adopted to preprocess the above datasets in the experiments.

Two aspects including classification accuracy and reduction rate are considered in this comparative experiment. Similarly, we perform experiments in the case of different ratios of unlabeled instances, where the ratio of unlabeled instances is set as 0.05 to 0.95 with the step of 0.05. CART, RBF-SVM and NB are used to evaluate the classification performance of different reduct sets. It's noted that since the feature subsets with best classification performance for these three classifiers are chosen respectively in NGAR, the obtained reduct sets may be different. Figs. 4-6 show the variations of classification accuracy with ratio of unlabeled data in CART, RBF-SVM and NB respectively. The classification accuracies of the original attribute sets are also given in the figures. Fig. 7 presents the variation of reduction rates with ratio of unlabeled data. The average classification accuracies and reduction rate of these four algorithms in different ratios of unlabeled instances are given in Tables 6 to 9 respectively.

From the above figures and tables, it can be seen that in terms of classification performance, all the above methods can find effective reduct sets. However, as shown in Tables 6 and 8, the average classification accuracies of NGAR are higher. Moreover, we can see that NGAR has a distinct advantage over the other three methods when RBF-SVM and NB are used as the classifier. In terms of reduction rate, NGAR has higher reduction rates. Namely, the reduct set obtained by NGAR owns less attributes and higher classification accuracy. Moreover, as to Semi-P, Semi-D, and MBAR, as the ratio of unlabeled data increases, the reduction rate turns worse. In contrast, the reduction rate of NGAR is more robust. The underlying reason is that they have different granular structures and evaluation indicators. As to Semi-P and Semi-D, the basic granular structure is equivalence class, which is only applicable to categorical data. Thus, data discretization is required for numerical data, which will unavoidably lead to the loss of information.

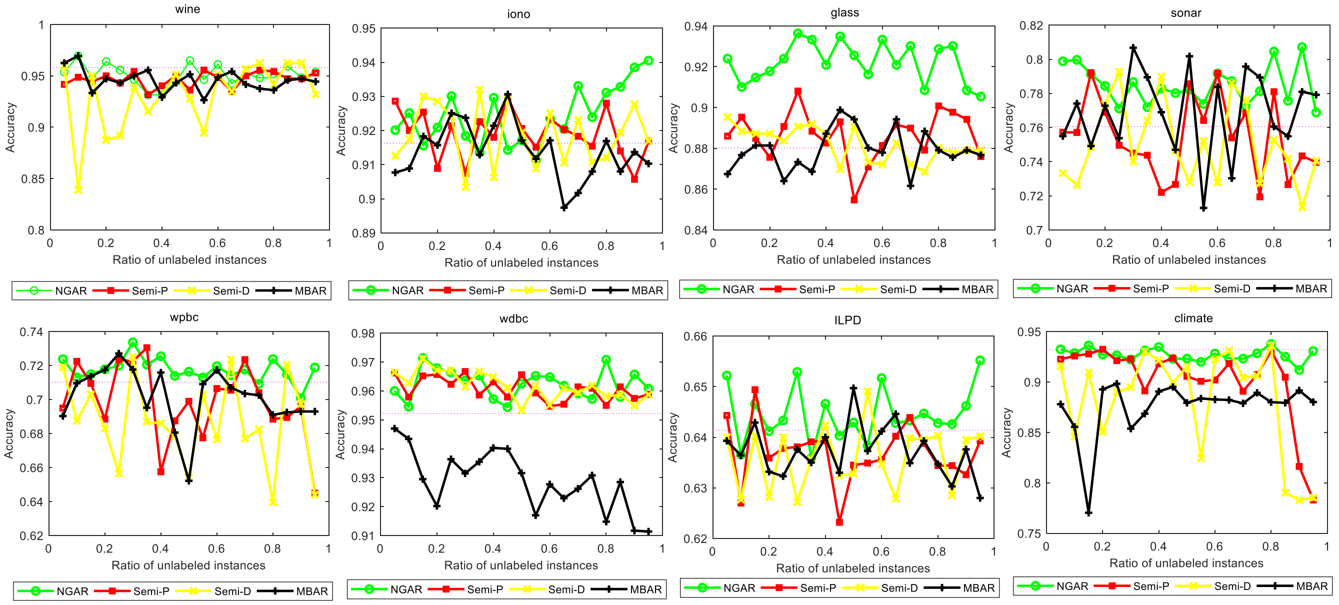

FIGURE 5. Comparisons of classification performance with existing reduction methods for partially labeled data (RBF-SVM).

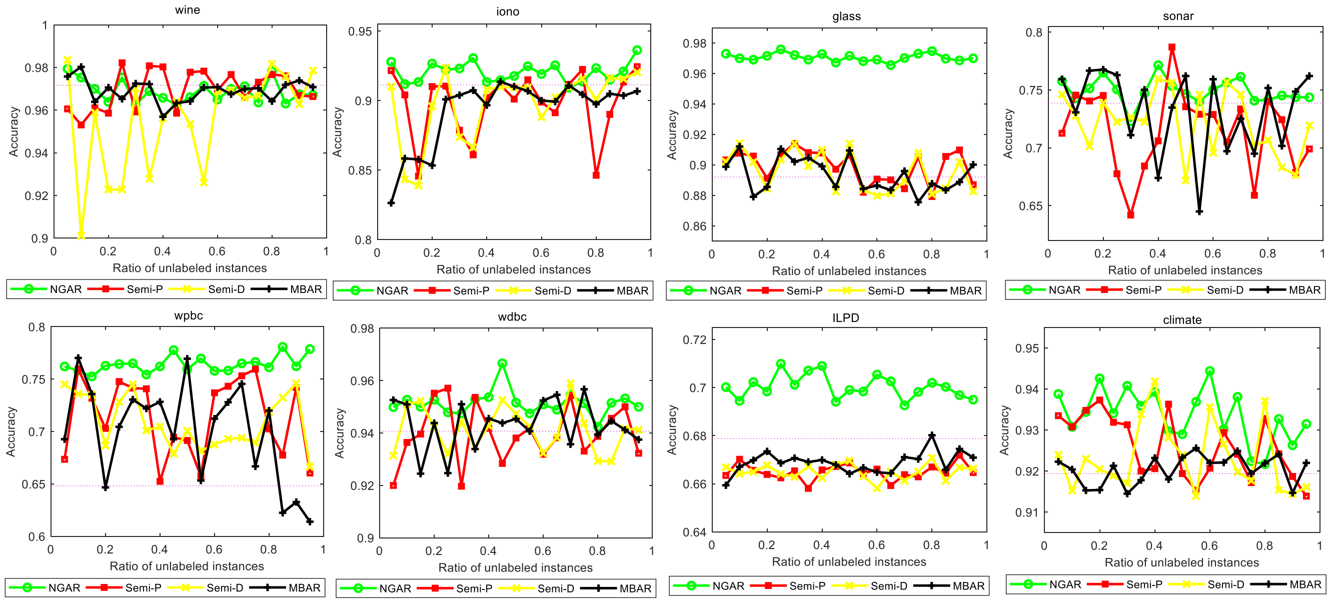

FIGURE 6. Comparisons of classification performance with existing reduction methods for partially labeled data (Naïve Bayes).

Whereas the neighborhood granulation is based on neighborhood relation, which can directly handle numerical data. The same problem can also be found in MBAR. Moreover, as to MBAR, each unlabeled data will be assigned with a pseudo-class symbol. Since each unlabeled data with pseudo-class symbol is required to be discernible, there may exist more redundant attributes in the obtained reduct set. In contrast, the significance introduced by us can measure the labeled and unlabeled parts simultaneously, which can fully use the original data without any change to it. Thus, our proposed method performs better.

C. THE INFLUENCE OF NEIGHBORHOOD THRESHOLD

In the framework of neighborhood rough set, the neighborhood threshold plays a major role in controlling the granularity levels of sample space. As mentioned above, different settings of neighborhood threshold will lead to the change of neighborhood granulation structure, which may further influence the significances of attribute sets. Accordingly, different reduct sets will be chosen by NGAR. Thus, in this subsection, we study the impact of neighborhood threshold on the performance of reduct set. Four datasets are considered in the experiments with three different percent

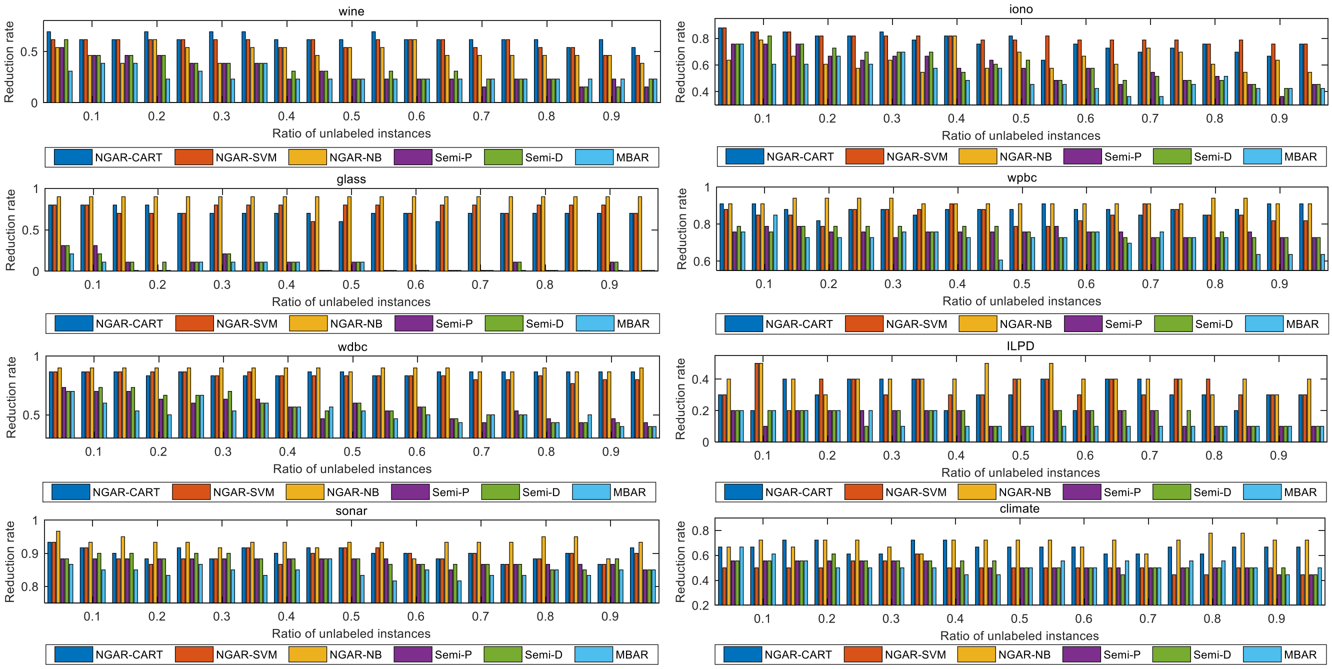

FIGURE 7. Comparisons of reduction rate with existing reduction methods for partially labeled data.

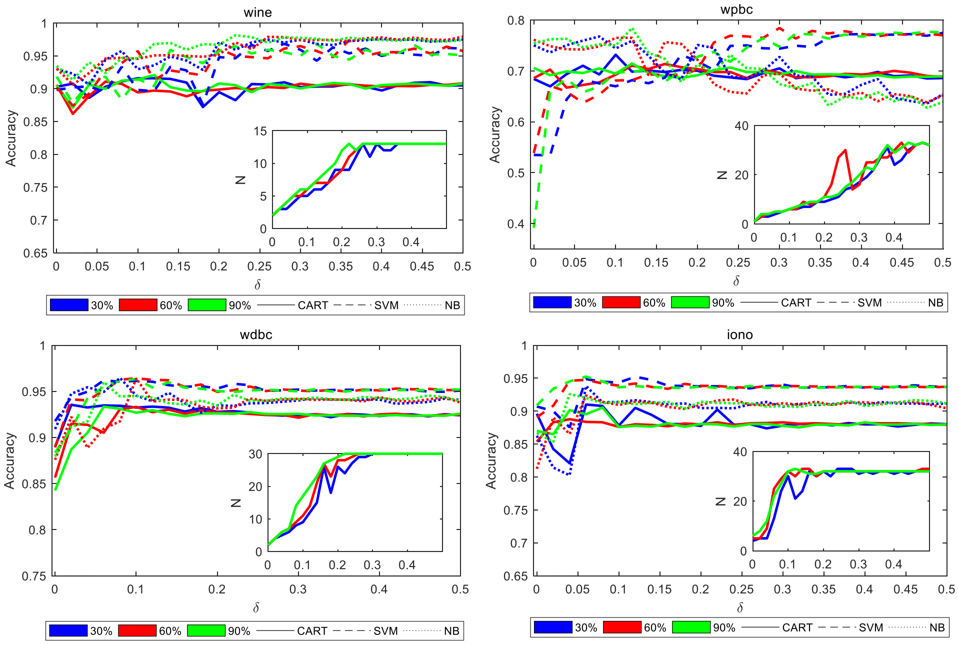

FIGURE 8. Variations of classification accuracies and feature numbers of reduct set with neighborhood threshold.

of unlabeled data, where the ratio of unlabeled data is set as 30%, 60%, and 90% respectively. In each ratio of unlabeled data, the neighborhood threshold δ ranges from 0 to 0.5 with step 0.02. Similarly, CART, RBF-SVM and NB are utilized to evaluate the classification accuracies of the obtained reduct sets. In the meanwhile, we report the number **N** of features in the corresponding reduct set. The variations of classification performance and the number **N** of selected features with neighborhood threshold are shown in Fig. 8. Broadly speaking, the classification accuracies increase at first, and then keep stable. Whereas the number of selected features increases with the threshold overall. There is also some small difference among the results obtained by CART, RBF-SVM and NB. Rough speaking, we can find that

TABLE 6. Average classification performance (CART).

Data	NGAR	Semi-P	Semi-D	MBAR
Wine	**0.9140**	0.9095	0.9006	0.9059
Iono	**0.8910**	0.8798	0.8771	0.8701
Glass	**0.9818**	0.9817	0.9813	0.9775
Wpbc	**0.7220**	0.6948	0.6903	0.6862
Wdbc	**0.9362**	0.9300	0.9305	0.9265
ILPD	**0.6651**	0.6542	0.6556	0.6557
Sonar	**0.7322**	0.6955	0.6924	0.6966
Climate	**0.9179**	0.9050	0.8948	0.8991
average	**0.8450**	0.8313	0.8278	0.8272

TABLE 7. Average classification performance (RBF-SVM).

Data	NGAR	Semi-P	Semi-D	MBAR
Wine	**0.9509**	0.9466	0.9311	0.9456
Iono	**0.9232**	0.9189	0.9188	0.9140
Glass	**0.9223**	0.8864	0.8820	0.8793
Wpbc	**0.7175**	0.6985	0.6865	0.7015
Wdbc	**0.9625**	0.9607	0.9617	0.9288
ILPD	**0.6445**	0.6370	0.6361	0.6372
Sonar	**0.7849**	0.7546	0.7506	0.7688
Climate	**0.9271**	0.9024	0.8819	0.8753
average	**0.8541**	0.8381	0.8311	0.8313

TABLE 8. Average classification performance (NB).

Data	NGAR	Semi-P	Semi-D	MBAR
Wine	0.9688	**0.9696**	0.9556	0.9692
Iono	**0.9205**	0.8979	0.8978	0.8928
Glass	**0.9706**	0.8991	0.8957	0.8933
Wpbc	**0.7642**	0.7142	0.7089	0.6994
Wdbc	**0.9514**	0.9398	0.9415	0.9431
ILPD	**0.7004**	0.6650	0.6651	0.6689
Sonar	**0.7490**	0.7144	0.7213	0.7318
Climate	**0.9337**	0.9259	0.9233	0.9204
average	**0.8698**	0.8407	0.8387	0.8399

TABLE 9. Average reduction rate.

Data	NGAR	NGAR	NGAR	Semi-	Semi-	MBA
	-C	-S	-N	P	D	R
Wine	**0.6275**	0.5789	0.4899	0.2995	0.3198	0.2632
Iono	0.7719	**0.8038**	0.6396	0.5805	0.5949	0.5199
Glass	0.7105	0.7579	**0.9000**	0.0889	0.0889	0.0521
Wpbc	0.8788	0.8501	**0.9187**	0.7528	0.7576	0.7193
Wdbc	0.8561	0.8333	**0.8912**	0.5579	0.5667	0.5228
ILPD	0.3158	0.3474	**0.4000**	0.1579	0.1632	0.1316
Sonar	0.9000	0.8939	**0.9307**	0.8746	0.8772	0.8447
Climate	0.6637	0.5029	**0.7047**	0.5146	0.5234	0.5205
average	0.7155	0.6960	**0.7343**	0.4783	0.4865	0.4468

[0.05,0.15] is an ideal interval for threshold, when we use CART and NB to evaluate the classification performance. As to RBF-SVM, [0.15, 0.3] is an appropriate interval for the neighborhood threshold to find a relatively optimal reduct set with minimum number of features and maximum classification accuracy.

V. CONCLUSIONS

Rough set theory is an important feature selection technique. However, most existing researches on it mainly focus on labeled data. To overcome such problem, in this paper, we present a new perspective to handle feature selection issue of partially labeled data. Firstly, we introduce several neighborhood granulation measures including neighborhood information granulation, neighborhood combination information granulation, and conditional neighborhood information granulation, which can measure the granular structure and knowledge implied in an information system or a decision system. Moreover, their properties are discussed systematically. Then, a novel concept of significance is proposed to measure the discernibility power of attributes in partially labeled data. Based on that, a feature selection method for partially labeled data is proposed. Finally, we conduct a series of experiments to verify the validity of our proposed method. The comparative results and analysis show the applicability and effectiveness of our proposed method. It's noted that in reality, there may exist both categorical and numerical features in the partially labeled data. Thus, in the future, our group will mainly investigate the mixed feature selection problem for partially labeled data.

REFERENCES

[1] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 619–632, Mar. 2013.

[2] Q. Hu, L. Zhang, Y. Zhou, and W. Pedrycz, "Large-scale multimodality attribute reduction with multi-kernel fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 1, pp. 226–238, Feb. 2018.

[3] C. Wang *et al.*, "A fitting model for feature selection with fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 4, pp. 741–753, Aug. 2017.

[4] Q. Hu, L. Zhang, D. Chen, W. Pedrycz, and D. Yu, "Gaussian kernel based fuzzy rough sets: Model, uncertainty measures and applications," *Int. J. Approx. Reasoning*, vol. 41, no. 4, pp. 453–471, 2010.

[5] Q. Hu, D. Yu, J. Liu, and C. Wu, "Neighborhood rough set based heterogeneous feature subset selection," *Inf. Sci.*, vol. 178, no. 18, pp. 3577–3594, 2008.

[6] X. Gu, Y. Li, and J. Jia, "Feature selection for transient stability assessment based on kernelized fuzzy rough sets and memetic algorithm," *Int. J. Elect. Power Energy Syst.*, vol. 64, pp. 664–670, Jan. 2015.

[7] S. Chebrolu and S. G. Sanjeevi, "Attribute reduction on real-valued data in rough set theory using hybrid artificial bee colony: Extended FTSBPSD algorithm," *Soft Comput.*, vol. 21, no. 24, pp. 7543–7569, 2017.

[8] D. Zouache and F. Ben Abdelaziz, "A cooperative swarm intelligence algorithm based on quantum-inspired and rough sets for feature selection," *Comput. Ind. Eng.*, vol. 115, pp. 26–36, Jan. 2018.

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

[10] Q. Sun, C. Wang, Z. Wang, and X. Liu, "A fault diagnosis method of smart grid based on rough sets combined with genetic algorithm and tabu search," *Neural Comput. Appl.*, vol. 23, nos. 7–8, pp. 2023–2029, 2013.

[11] Z. Pawlak, "Rough sets," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 145–172, 1982.

[12] H. Zhao and K. Qin, "Mixed feature selection in incomplete decision table," *Knowl.-Based Syst.*, vol. 57, pp. 181–190, Feb. 2014.

[13] Y. Y. Yao, "Relational interpretations of neighborhood operators and rough set approximation operators," *Inf. Sci.*, vol. 111, nos. 1–4, pp. 239–259, 1998.

[14] X. Fan, W. Zhao, C. Wang, and Y. Huang, "Attribute reduction based on max-decision neighborhood rough set model," *Knowl.-Based Syst.*, vol. 151, pp. 16–23, Jul. 2018.

[15] Y. Chen, Z. Zeng, and J. Liu, "Neighborhood rough set reduction with fish swarm algorithm," *Soft Comput.*, vol. 21, no. 23, pp. 6907–6918, 2017.

[16] H. Chen, T. Li, Y. Cai, C. Luo, and H. Fujita, "Parallel attribute reduction in dominance-based neighborhood rough set," *Inf. Sci.*, vol. 373, pp. 351–368, Dec. 2018.

[17] T. Zeng and L. Zhu, "Uncertainty measures of neighborhood system-based rough sets," *Knowl.-Based Syst.*, vol. 86, pp. 57–65, Sep. 2015.

[18] A. Sun, Y. Liu, and E.-P. Lim, "Web classification of conceptual entities using co-training," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 14367–14375, 2011.

[19] Y. Liu and K. Kirchhoff, "Graph-based semisupervised learning for acoustic modeling in automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 1946–1956, Nov. 2016.

[20] Y. Liu, K. Wen, Q. Gao, X. Gao, and F. Nie, "SVM based multi-label learning with missing labels for image annotation," *Pattern Recognit.*, vol. 78, pp. 307–317, Jun. 2018.

[21] D. Miao, C. Gao, N. Zhang, and Z. Zhang, "Diverse reduct subspaces based co-training for partially labeled data," *Int. J. Approx. Reason.*, vol. 52, no. 8, pp. 1103–1117, 2011.

[22] J. Dai, Q. Hu, J. Zhang, H. Hu, and N. Zheng, "Attribute selection for partially labeled categorical data by rough set approach," *IEEE Trans. Cybern.*, vol. 47, no. 9, pp. 2460–2471, Sep. 2017.

[23] W. Zhang, D. Miao, C. Gao, and X. Yue, "Co-training based attribute reduction for partially labeled data," in *Rough Sets and Knowledge Technology* (Lecture Notes in Computer Science), vol. 8818. Berlin, Germany: Springer-Verlag, 2014, pp. 77–88.

[24] J. Ren, Z. Qiu, W. Fan, H. Cheng, and P. Yu, "Forward semisupervised feature selection," in *Proc. Adv. Knowl. Discovery Data Mining*, Osaka, Japan, 2008, pp. 970–976.

[25] P. Zhu, Q. Xu, Q. Hu, C. Zhang, and H. Zhang, "Multi-label feature selection with missing labels," *Pattern Recognit.*, vol. 74, pp. 488–502, Feb. 2018.

[26] J. Liang and Y. Qian, "Information granules and entropy theory in information systems," *Sci. China F-Inf. Sci.*, vol. 51, no. 10, pp. 1427–1444, 2008.

[27] J. Liang and Z. Shi, "The information entropy, rough entropy and knowledge granulation in rough set theory," *Int. J. Uncertainty Fuzziness Knowl.-Based Syst.*, vol. 12, no. 1, pp. 37–46, 2004.

[28] Y. Qian and J. Liang, "Combination entropy and combination granulation in incomplete information system," in *Rough Sets and Knowledge Technology* (Lecture Notes in Artificial Intelligence), vol. 4062. Berlin, Germany: Springer-Verlag, 2006, pp. 184–190.

[29] Y. Chen, K. Wu, X. Chen, C. Tang, and Q. Zhu, "An entropy-based uncertainty measurement approach in neighborhood systems," *Inf. Sci.*, vol. 279, pp. 239–250, Sep. 2014.

[30] (2005). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[31] Q. Hu, D. Yu, and Z. Me, "Neighborhood classifiers," *Expert Syst. Appl.*, vol. 34, no. 2, pp. 866–876, 2008.

BINGYANG LI was born in Jiangsu, china, in 1993. He received the B.S. degree in automation from the Logistics Engineering College, Shanghai Maritime University, Shanghai, China, in 2015, where he is currently pursuing the Ph.D. degree. His current research interests include rough set theory and granular computing.

JIANMEI XIAO received the M.S. degree from Dalian Maritime University, Dalian, China, in 1988. She is currently a Full Professor with the Department of Electrical Engineering, Shanghai Maritime University. Her current research interests include rough set theory, intelligent control, and intelligent decision making.

XIHUAI WANG received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China. He is currently a Full Professor with the Department of Electrical Engineering, Shanghai Maritime University. His current research interests include fuzzy set theory, rough set theory, approximate reasoning, and intelligent information processing.