Local Feature Selection for Large-scale Data Sets with Limited Labels

Tian Yang, Yanfang Deng, Bin Yu, Yuhua Qian*, Jianhua Dai*

Abstract—Processing large-scale data sets with limited labels has always been a difficult task in data mining. Facing this difficulty, two local feature selection algorithms, LARD and LRSD, have been proposed based on dependency degree, which can process partially labeled data sets and greatly improve the computational efficiency. However, it is very difficult for these algorithms to calculate large-scale data with millions of samples on a typical personal computer. Although the related family method is a more efficient approach than dependency degree, it cannot be used for partially labeled large-scale data. As a result, a local feature selection method based on related family is proposed to accelerate data processing in the paper. Experiments show that the proposed algorithm can run 405 times faster than LARD on partially labeled data sets and maintain high classification accuracy. In addition, this new algorithm can effectively process partially labeled large-scale data sets with 5,000,000 samples or 20,000 features on a typical personal computer.

Index Terms—Data mining, Semi-supervised learning, Local feature selection, Rough set, Related family

1 INTRODUCTION

With the widespread use of various sensors, the rate of data generation in human society has become much higher than the increase in computing power. At the same time, the large amount of data lacks labels for sample collection, which is both time-consuming and costly. Therefore, the study of big data not only involves large-scale samples and ultra-high dimensions, but also may encounter problems such as the lack of labels, which has become a hot topic in the field of artificial intelligence and data mining in recent years [1-12]. To process data with limited labels, a plenty of semi-supervised learning methods have been proposed, such as co-training [1], semi-supervised support vector machines [2], label propagation [3], graph-based semi-supervised learning method [4], etc. Hou et al. [5] put forward a one-pass method to learn and simultaneously evolve instances from data with incremental and decremental features, and then a safe classification algorithm [6]. Wang et al. [8] proposed a scalable graph-based semi-supervised learning algorithm called Efficient Anchor Graph Regularization, and then an efficient Hierarchical Anchor Graph Regularization approach [9]. And Yu et al. [10-12] brought up the progressive semi-supervised ensemble learning approach. Although these algorithms have achieved remarkable performance in partially labeled data processing, their efficiency needs to be improved. Hence, how to develop more efficient large-scale data processing methods with limited computing resources remains a problem in data mining.

Feature selection is an effective method for large-scale data processing, which can not only reduce the dimensions of data, thus reducing the computational cost and avoiding the “curse of dimension", but also remove the noisy data and improve the accuracy and generalization ability of machine learning. Feature selection methods can be classified into three categories[13-17]: filter, wrapper and embedded. Neural networks can effectively perform feature selection[18-20], but the features extracted by neural networks are generally poorly interpretable and susceptible to noise interference. Therefore, more and more scholars advocate the study of interpretable and more robust data processing methods. Granular computing [21-24] has been widely used in artificial intelligence, data mining and intelligent decision making [24-29] due to its security, robustness and interpretability.

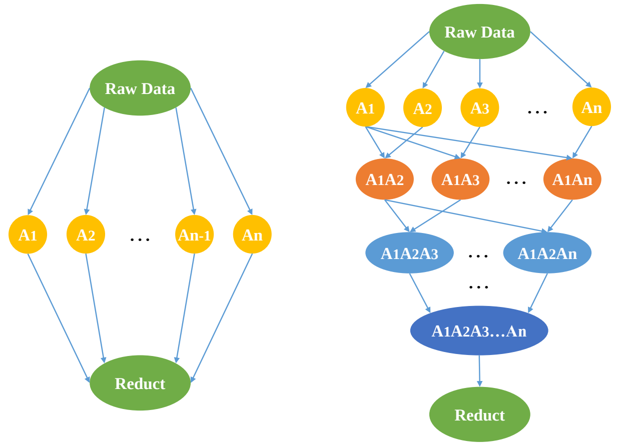
Rough set [22, 23] is a typical granular computing model and an important theoretical branch of machine learning and data mining. Feature selection (also known as attribute reduction) algorithm based on rough set has become a research hotspot in recent years [30-42]. Hu et al. [43] introduced the neighborhood rough set (NRS) and designed the feature selection algorithm. Qian et al. [44] proposed a positive region acceleration method, which provides an accelerated method for feature selection. Dai et al. [45] introduced the gain ratio into the fuzzy rough set theory. Xia et al. [33] innovated the neighborhood generation method, and proposed a self-adaptive feature selection algorithm based on ball neighborhood, which has linear time complexity relative to the sample size. For partially labeled data processing, scholars have proposed several strategies[46-49]. Dai et al. [46] proposed two semi-supervised feature selection algorithms based on discernibility pairs. Liu et al. [47] added false labels to

- *T. Yang, Y. Deng, B. Yu, J. Dai are with the Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing (2018TP1018), Innovation and Entrepreneurship Training Program of Hunan Xiangjiang Artificial Intelligence Academy, Hunan Normal University, Changsha 420081, China. E-mail: math_yangtian@126.com, dengyanfang21@126.com, yubin20070119@sina.com, jhdai@hunnu.edu.cn.*
- *Y. Qian is with the Institute of Big Data Science and Industry, Shanxi University, Taiyuan 03006, China. E-mail: jinchengqyh@126.com.*
- *Corresponding authors: Yuhua Qian and Jianhua Dai.*

unlabeled samples and proposed a semi-supervised learning method. Although these algorithms have achieved prominent performance in feature selection, due to their high complexity in time or space, none of them can compute million-scale samples or ten-thousand-scale data on a typical personal computer. To deal with large-scale partially labeled data, it is necessary to explore new computing modes and design feature selection algorithms with low time/space complexity.

Two effective strategies are adopted to construct a new computing framework: (1) Localized computation. The first corresponding author, Qian [48] and his co-author initially proposed a local rough set model, in which only the information granules related to the target objects are computed, rather than all the information granules (as shown in Fig. 3.1). On this basis, a local feature selection algorithm (LRSD) was designed. Wang et al. [49] designed a local neighborhood feature selection algorithm (LARD) based on local neighborhood rough sets. LARD and LRSD can not only effectively process partially labeled data, but also greatly improve the computational efficiency. However, with the continuous expansion of data scale, the localized computation strategy still faces enormous challenges. (2) Efficient feature evaluation. In most granular computing models, multiple features are regarded as a feature subset to generate granules, and each feature is evaluated by the granule difference before and after the addition or removal of features in the subset. In this case, since the feature subset is constantly changing in the feature selection processing, the information granules need to be repeatedly computed. To simplify the processing, the first author, Yang and her co-author initially proposed the related family method[50]. Compared with other methods, the related family method only requires a single calculation based on each single feature to generate information granules (the computation process is shown in Fig. 1.1). Fujita and his co-author verified the effectiveness and efficiency of the related family method and proposed the incremental feature selection algorithms for the dynamic covering decision system based on related family in several literatures [51-53]. In a bid to improve the classification accuracy of noise data processing, Ou et al. [54] designed a feature selection algorithm for variable precision covering rough sets based on related family. Although the feature selection algorithm based on related family has high computational efficiency, it cannot directly process partially labeled data. To efficiently process partially labeled large-scale data, a local feature selection method based on related family was proposed in this paper.

Firstly, local upper and lower approximate operators based on covering rough sets are introduced; on this basis, a new local feature selection method based on related family, namely, the local related family, is proposed. Then, a heuristic local feature selection algorithm with linear time and space complexities is designed. Experiments show that the proposed algorithm can effectively process partially labeled data sets while maintaining the classification accuracy, and its running speed is hundreds of times faster than LARD [49] and GBNRS [33]. In addition, as the scale of the data sets increases, so does the efficiency advantage. As a result, the new algorithm can process large-scale data sets of 5 million samples or 20,000 features on personal computers, whereas other algorithms can only compute less than 10% of that scale.



(a) Related family (b) Dependency degree

Fig. 1.1. The computation process comparison of related family and dependency degree.

The contributions of this paper include: (1) a new feature evaluation framework for partially labeled data, namely, the local related family, is proposed; (2) an efficient feature selection algorithm for partially labeled data with linear time and space complexities is designed, and its feature selection speed is even increased by 405 times.

The rest of the paper is as follows. In section 2, the basic concepts of local rough sets and variable precision covering rough sets are introduced. In section 3, a new local feature selection method is proposed based on local rough sets and related family methods. In section 4, a local feature selection algorithm with linear time complexity is designed. The results of numerical experiments are analyzed in Section 5, and finally, a brief conclusion is given in Section 6.

2 Background Knowledge

2.1 Local Rough Set

Local rough set (LRS), proposed by Qian et al. [48], is an efficient approach for processing partially labeled data. Some basic notions in local rough sets are introduced in this study.

Let U be a nonempty finite set (called universe), $R \in U \times U$ be an equivalence relation. For $\forall w, z \in U$, if $(w,z) \in R$, then w is equivalent to z under the relation R, and $(z,w) \in R$. For $\forall w \in U$, the equivalence class of object w by equivalence relation R is

$$[w]_R = \{z \in U \mid (w,z) \in R\}.$$

Definition 1 [48]. Give an universe U and an equivalence relation $R \in U \times U$, for $\forall W \subseteq U$, the α-lower approximation operator $\underline{LR}_\alpha(W)$ and the β-upper approximation operator $\overline{LR}_\beta(W)$ of W are defined as

$$\underline{LR}_\alpha(W) = \{w \mid \theta(W/[w]_R) \ge \alpha,\ w \in W\} \tag{1}$$

$$\overline{LR}_\beta(W) = \cup\{[w]_R \mid \theta(W/[w]_R) > \beta,\ w \in W\} \tag{2}$$

where $0 < \beta \le \alpha \le 1$,

$$\theta(W/[w]_R) = \frac{|W \bigcap [w]_R|}{|[w]_R|}$$

is the inclusion degree of $[w]_R$ with respect to W, $|*|$ is the number of elements in $*$, $[w]_R$ is an equivalence class of object w by equivalence relation R. The $<\underline{LR}_\alpha, \overline{LR}_\beta>$ is called LRS.

2.2 Variable Precision Covering Rough Set

Covering rough set (CRS) [55, 56] is an important extension of classical rough set. To improve the effect of noise data processing, CRS is extended to variable precision covering rough set (VPCRS) [57, 58].

Let $\mathbb{C}$ be a non-empty subset family of universe U, if $U = \bigcup \mathbb{C}$, the $\mathbb{C}$ is a covering of universe U.

Definition 2 [58]. Let $\mathbb{C}$ be a covering of universe U, for $\forall W \subseteq U$, the α-lower and α-upper approximation operators are defined by

$$\underline{CM}^{\alpha}_{\mathbb{C}}(W) = \bigcup\{G \in \mathbb{C} \mid \theta(W/G) \geq \alpha\} \quad (3)$$

$$\overline{CM}^{\alpha}_{\mathbb{C}}(W) = \bigcup\{G \in \mathbb{C} \mid \theta(W/G) > 1-\alpha\} \quad (4)$$

where $0.5 < \alpha \leq 1$. The subscript $\mathbb{C}$ can be omitted if there is no confusion.

3 Local Feature selection

To improve the computational efficiency, a local feature selection method based on related family is studied in this paper. Firstly, the local upper and lower approximation operators based on VPCRS are introduced, and the local reduction is defined based on this model. Next, a local feature selection method is initially proposed.

According to Definition 2, it can be seen that the α-lower approximation operator of W may include samples other than W. In order to include the lower approximation of W in W, upper and lower approximation operators are reconstructed, as shown below:

Definition 3. Let $\mathbb{C}$ be a covering of universe U, for $\forall W \subseteq U$, the local upper and lower approximation operators are defined by

$$\underline{LC}^{\alpha}_{\mathbb{C}}(W) = \bigcup\{G \bigcap W \mid \theta(W/G) \geq \alpha,\ G \in \mathbb{C}\} \quad (5)$$

$$\overline{LC}^{\beta}_{\mathbb{C}}(W) = \bigcup\{G \mid \theta(W/G) \geq \beta,\ G \in \mathbb{C}\} \quad (6)$$

where $0 < \beta \leq 0.5 < \alpha \leq 1$. The subscript $\mathbb{C}$ can be omitted if there is no confusion.

Let U be a universe, CF be a covering family on U, D be the decision attribute, (U, CF, D) is called a covering decision information system (CDIS).

Since $\bigcup CF$ is still a covering on U, for any subset W of U, the local lower approximation on (U, CF, D) can be defined as

$$\underline{LC}^{\alpha}_{\bigcup CF}(W) = \bigcup\{G \bigcap W \mid \theta(W/G) \geq \alpha,\ G \in \bigcup CF\}.$$

In practice, collecting sample labels is both time-consuming and expensive, therefore a lot of data lack labels. In this paper, the set of all labeled samples is regarded as the target set.

Suppose $TS \subseteq U$ is a target set, then TS is the set of all labeled samples and $U - TS$ is the unlabeled sample set. The partition of the target set TS with respect to D is defined as

$$TS/D = \{W_i \mid W_i = D_i \bigcap TS,\ D_i \in U/D\}.$$

Since unlabeled samples can also provide some useful information, all samples (both labeled and unlabeled) are used in this paper to approximate the target set. Then the local positive region of D concerning target set TS is defined as

$$POS_{\bigcup CF}(TS) = \bigcup_{W \in TS/D} \underline{LC}^{\alpha}_{\bigcup CF}(W).$$

Proposition 1. Given a CDIS (U, CF, D), and the target set $TS \subseteq U$. If $H \subseteq K \subseteq CF$, the $POS_{\bigcup H}(TS) \subseteq POS_{\bigcup K}(TS)$.

Proof: Since $POS_{\bigcup K}(TS) = \bigcup_{W \in TS/D} \underline{LC}^{\alpha}_{\bigcup K}(W)$, $\underline{LC}^{\alpha}_{\bigcup K}(W) = \bigcup\{G \bigcap W \mid \theta(W/G) \geq \alpha,\ G \in \bigcup K\}$.

Then $POS_{\bigcup K}(TS) = \bigcup\{\underline{LC}^{\alpha}_{\bigcup K}(W) \mid W \in TS/D\}$
$= \bigcup\{G \bigcap W | G \in \bigcup K \text{ and } \exists W \in TS/D \text{ s.t. } \theta(W/G) \geq \alpha\}$
$= \bigcup\{G \bigcap W | G \in \bigcup H \text{ and } \exists W \in TS/D \text{ s.t. } \theta(W/G) \geq \alpha\}$
$\bigcup(\bigcup\{G \bigcap W | G \in \bigcup(K-H) \text{ and } \exists W \in TS/D \text{ s.t. } \theta(W/G) \geq \alpha\})$
$= POS_{\bigcup H}(TS) \bigcup (\bigcup\{G \bigcap W | G \in \bigcup(K-H) \text{ and } \exists W \in TS/D$
$\text{s.t. } \theta(W/G) \geq \alpha\})$. Thus $POS_{\bigcup H}(TS) \subseteq POS_{\bigcup K}(TS)$. □

The purpose of local feature selection for CDIS is to find a minimal feature subset (called a local feature reduct or a local attribute reduct in this paper) maintaining the local positive region unchanged.

Definition 4. Given a CDIS (U, CF, D) and the target set $TS \subseteq U$. For $\mathbb{C} \in CF$, if $POS_{\bigcup CF}(TS) = POS_{\bigcup(CF - \{\mathbb{C}\})}(TS)$, then $\mathbb{C}$ is local redundancy in CF with respect to TS, otherwise $\mathbb{C}$ is local necessary in CF with respect to TS. For $K \subseteq CF$, if each feature $\mathbb{C} \in K$ is local necessary and $POS_{\bigcup CF}(TS) = POS_{\bigcup K}(TS)$, then K is called a local feature reduct (or a local attribute reduct, short for a local reduct) of CF with respect to TS.

Let $RED(CF)$ be the set of all local reducts of CF with respect to TS, then the core of local reduct is $CORE(CF) = \bigcap RED(CF)$.

Then a new local feature selection method, named local related family is developed to compute all local reducts.

The α-consistent set is a key notion in local related family method, composed of all α-consistent information granules whose inclusion degree for a decision class is greater than α, defined as following.

Definition 5. Given a CDIS (U, CF, D) and the target set $TS \subseteq U$, $TS/D = \{W_1, W_2, \ldots, W_s\}$ is the partition of the target set TS with respect to D. Then α-consistent set of TS is defined as (7):

$$M^{\alpha}_{\bigcup CF}(TS) = \{G \in \bigcup CF \mid \exists W_i \in TS/D \text{ s.t. } \theta(W_i/G) \geq \alpha\} \quad (7)$$

Each element in the α-consistent set is called an α-consistent information granule of CF with respect to TS. If $w_i \in G$ and the label of w_i is the same as the label of most samples in G, w_i is consistently included by G, and noted as $w_i \underline{\in} G$.

In an α-consistent information granule, there may be unlabeled samples and samples with different labels, thus the consistent inclusion relation is defined in Definition 5. For any sample $w_i \in POS_{\bigcup CF}(TS)$, there is at least one feature that generates an α-consistent information granule G consistently including sample w_i.

Since the local positive region does not contain any unlabeled samples, we only need to calculate the related sets of labeled samples instead of all samples, which saves a lot of time (as shown in Fig. 3.1). The local related family is defined as follows:

Definition 6. Given a CDIS (U, CF, D) and the target set $TS \subseteq U$. For $\forall w_i \in TS$, the local related set of w_i is defined as (8), and the local related family of target set TS is defined as (9):

$$\gamma(w_i) = \{\mathbb{C} \in CF \mid \exists G \in M_{\cup CF}^{\alpha}(TS) \text{ s.t. } w_i \underline{\in} G \in \mathbb{C}\} \tag{8}$$

$$LR(TS, CF, D) = \{\gamma(w_i) \mid \gamma(w_i) \neq \phi,\ w_i \in TS\} \tag{9}$$

where $M_{\cup CF}^{\alpha}(TS)$ is the α-consistent set of TS.

Each local related set $\gamma(w_i)$ in the local related family is the collection of all features which generate at least one α-consistent information granule G consistently including sample w_i. If $\gamma(w_i) \neq \phi$, then the sample w_i belongs to the local positive region. Where each information granule is generated based on a single feature without the need for multiple calculations.

Unlike the related family, the local related family method does not need to calculate the relevant information of all samples (no need to compute the related sets r_{n+1}, r_{n+2}, …, r_{n+m}), but only calculates the samples in the target set (only calculate related sets r_1, r_2, …, r_n), which greatly reduces the time consumption. As shown in Fig. 3.1.

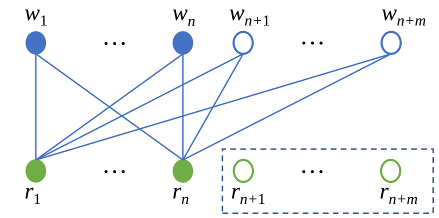


Fig. 3.1. The local related family (no need to compute the related sets r_{n+1}, r_{n+2}, …, r_{n+m}).

Theorem 1. H is a local reduct of CF if and only if H is a minimal feature subset that satisfying the following condition: $H \cap \gamma(w_i) \neq \phi$ for every $\gamma(w_i) \in LR(TS, CF, D)$.

Proof: ($\Rightarrow$) Suppose H is a local reduct of CF, then $POS_{\cup CF}(TS) = POS_{\cup H}(TS)$. For $\forall \gamma(w_i) \in LR(TS, CF, D)$, since $\gamma(w_i)$ is nonempty, we have $w_i \in POS_{\cup CF}(TS)$ and $w_i \in POS_{\cup H}(TS)$. It implies there exists $G \in M_{\cup H}^{\alpha}(TS)$ and $\mathbb{C} \in H$ such that $w_i \underline{\in} G \in \mathbb{C}$. It is evident that $G \in M_{\cup CF}^{\alpha}(TS)$ and $\mathbb{C} \in \gamma(w_i)$. Therefore $\mathbb{C} \in \gamma(w_i) \cap H$, then $H \cap \gamma(w_i) \neq \phi$. Since each feature $\mathbb{C} \in H$ is local necessary in H, H is a minimal feature subset that satisfying the condition.

($\Leftarrow$) Suppose H is a minimal feature subset that satisfying the condition. For $\forall w_i \in POS_{\cup CF}(TS)$, since $\gamma(w_i) \neq \phi$ and $H \cap \gamma(w_i) \neq \phi$, suppose $\mathbb{C} \in H \cap \gamma(w_i)$, then there is $G \in \mathbb{C}$ such that $w_i \underline{\in} G \in M_{\cup H}^{\alpha}(TS)$. Then $w_i \in POS_{\cup H}(TS)$ and $POS_{\cup CF}(TS) \subseteq POS_{\cup H}(TS)$. Since $H \subseteq CF$, then $POS_{\cup H}(TS) \subseteq POS_{\cup CF}(TS)$. Thus $POS_{\cup H}(TS) = POS_{\cup CF}(TS)$. Furthermore, H is a minimal feature subset that satisfying the condition, which means each feature $\mathbb{C} \in H$ is local necessary in H. Therefore H is a local reduct of CF. □

Proposition 2. The $CORE(CF) = \{\mathbb{C} \mid \exists w_i \in POS_{\cup CF}(TS)$ s.t. $\gamma(w_i) = \{\mathbb{C}\}\}$.

Proof: ($\Rightarrow$) Suppose $\mathbb{C} \in CORE(CF)$, then $POS_{\cup CF}(TS) \supset POS_{\cup(CF-\{\mathbb{C}\})}(TS)$. Let $w_i \in POS_{\cup CF}(TS) - POS_{\cup(CF-\{\mathbb{C}\})}(TS)$, then there exists a α-consistent information granule G such that $w_i \underline{\in} G \in \mathbb{C}$, and $\mathbb{C}$ is the only feature which generates at least one α-consistent information granule G consistently including sample w_i, thus $\gamma(w_i) = \{\mathbb{C}\}$.

($\Leftarrow$) Suppose $\gamma(w_i) = \{\mathbb{C}\}$, then $\mathbb{C}$ is the only feature which generates at least one α-consistent information granule G consistently including sample w_i. Therefore $w_i \notin POS_{\cup(CF-\{\mathbb{C}\})}(TS)$, which means $\mathbb{C} \in CORE(CF)$. □

Based on the local related family, a new feature selection method is proposed based on Boolean functions.

Definition 7. Let (U, CF, D) be a CDIS, $TS \subseteq U$ be the target set, and the covering family $CF = \{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_k\}$. The local related function $f(TS, CF, D)$ is defined as:

$$f(TS, CF, D)(\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_k) = \wedge\{(\vee\gamma(w_i)) \mid \gamma(w_i) \in LR(TS, CF, D)\} \tag{10}$$

Where $LR(TS, CF, D)$ is the local related family of target set TS, k Boolean variables $\mathbb{C}_1$, $\mathbb{C}_2$, …, $\mathbb{C}_k$ correspond to k coverings $\mathcal{C}_1$, $\mathcal{C}_2$, …, $\mathcal{C}_k$.

Theorem 2. Given a CDIS (U, CF, D) and the target set $TS \subseteq U$. $LR(TS, CF, D)$ is the local related family of target set TS, $f(TS, CF, D)$ is the local related function. If the $g(TS, CF, D) = (\wedge P_1) \vee (P_2) \vee \ldots \vee (\wedge P_y)$ is the simplified disjunctive form derived from $f(TS, CF, D)$ by using the conjunction and disjunction rules, and for $j = 1, 2, \ldots, y$, every element in P_j is unique. Then $RED(CF) = \{P_1, P_2, \ldots, P_y\}$.

Proof: ($\Leftarrow$) For $j = 1, 2, \ldots, y$ and $\forall \gamma(w_i) \in LR(TS, CF, D)$, since $\wedge P_j \leq \vee\gamma(w_i)$, then $P_j \cap \gamma(w_i) \neq \phi$. Let $P_j^1 = P_j - \{\mathbb{C}\}$ for any $\mathbb{C} \in P_j$, then $g(TS, CF, D) < \vee_{t=1}^{j-1}(\wedge P_t) \vee (\wedge P_j^1) \vee (\vee_{t=j+1}^{y}(P_t))$. Suppose $P_j^1 \cap \gamma(w_i) \neq \phi$ for every $\gamma(w_i) \in LR(TS, CF, D)$, then $\wedge P_j^1 \leq \vee\gamma(w_i)$. That means $g(TS, CF, D) \geq \vee_{t=1}^{j}(\wedge P_t) \vee (\wedge P_j^1) \vee (\vee_{t=j+1}^{y}(\wedge P_t))$, which is a contradiction. Therefore, $\exists \gamma(w_i) \in LR(TS, CF, D)$ such that $P_j^1 \cap \gamma(w_i) = \phi$. Thus P_j is the local reduct of CF.

($\Rightarrow$) For $\forall R \in RED(CF)$ and $\forall \gamma(w_i) \in LR(TS, CF, D)$, since $R \cap \gamma(w_i) \neq \phi$, thus $f(TS, CF, D) \wedge (\wedge R) = \wedge(\vee\gamma(w_i)) \wedge (\wedge R) = \wedge R$. That means $\wedge R \leq f(TS, CF, D) = g(TS, CF, D)$. Suppose for $j = 1, 2, \ldots, y$, $P_j - R \neq \phi$, then $\exists \mathbb{C}_j \in P_j - R$. By rewriting $g(TS, CF, D) = (\vee_{j=1}^{y}\mathbb{C}_j) \wedge \Phi$, then $\wedge R \leq \vee_{j=1}^{y}\mathbb{C}_j$. Therefore $\exists \mathbb{C}_j$ such that $\wedge R \leq \mathbb{C}_j$, it implies there is $\mathbb{C}_j \in R$, which is a contradiction. That means $\exists P_j$ such that $P_j - R = \phi$, that is $P_j \subseteq R$, since R and P_j are both the local reduct, it means $P_j = R$. Thus $RED(CF) = \{P_1, P_2, \ldots, P_y\}$. □

An example is given to further illustrate the reduction process of this method presented in this paper.

Example 1. Given a CDIS (U, CF, D), where $U = \{w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8\}$, $CF = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4\}$, $D = \{1, 1, 1, 2, *, *, *, *\}$, where * means no label. Let parameter α=0.65 and target set $TS = \{w_1, w_2, w_3, w_4\}$.

$$\mathcal{C}_1 = \{\{w_1, w_2, w_3\}, \{w_2, w_7, w_8\}, \{w_3, w_5\}, \{w_4, w_6\}\},$$

$$\mathcal{C}_2 = \{\{w_1, w_2\}, \{w_2, w_5, w_6, w_8\}, \{w_3, w_7\}, \{w_4\}\},$$

$$\mathcal{C}_3 = \{\{w_1, w_3\}, \{w_2, w_4\}, \{w_3, w_7, w_6, w_8\}, \{w_4, w_5\}\},$$

$$\mathcal{C}_4 = \{\{w_1, w_3, w_5\}, \{w_2, w_3, w_4, w_5\}, \{w_2, w_3, w_4, w_5, w_6\}, \{w_4, w_5, w_6, w_7, w_8\}\}.$$

$$TS / D = \{W_1, W_2\} = \{\{w_1, w_2, w_3\}, \{w_4\}\}.$$

Then we construct the α-consistent set and the local positive region. Since $\{w_1, w_2, w_3\} \subseteq W_1$, the inclusion degree $\theta(W_1 / \{w_1, w_2, w_3\}) = 1 > 0.65$, then information granule

$\{w_1, w_2, w_3\}$ is α-consistent. Therefore, $\{w_1, w_2, w_3\}$ is added to the α-consistent set. Similarly, the inclusion degrees of other information granules relative to W_1 or W_2 are calculated respectively to get the α-consistent set:

$$M^{\alpha}_{\bigcup CF}(TS) = \{\{w_1, w_2, w_3\}, \{w_1, w_2\}, \{w_4\}, \{w_1, w_3\}, \{w_1, w_3, w_5\}\}.$$

$$POS_{\bigcup CF}(TS) = \{w_1, w_2, w_3, w_4\}.$$

Next, the local related sets of all $w_i \in TS$ are obtained based on the α-consistent set and the local positive region. For a labeled sample w_1, since there are four α-consistent information granules $\{w_1, w_2, w_3\} \in \mathcal{C}_1$, $\{w_1, w_2\} \in \mathcal{C}_2$, $\{w_1, w_3\} \in \mathcal{C}_3$, and $\{w_1, w_3, w_5\} \in \mathcal{C}_4$ which contain w_1, then the local related set of w_1 is $\gamma(w_1) = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4\}$. Similarly, we can obtain other local related sets:

$$\gamma(w_1) = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4\}, \quad \gamma(w_2) = \{\mathcal{C}_1, \mathcal{C}_2\},$$
$$\gamma(w_3) = \{\mathcal{C}_1, \mathcal{C}_3, \mathcal{C}_4\}, \quad \gamma(w_4) = \{\mathcal{C}_2\}.$$

As a result, the local related family is obtained:

$$LR(TS, CF, D) = \{\gamma(w_1), \gamma(w_2), \gamma(w_3), \gamma(w_4)\} = \{\{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4\}, \{\mathcal{C}_1, \mathcal{C}_2\}, \{\mathcal{C}_1, \mathcal{C}_3, \mathcal{C}_4\}, \{\mathcal{C}_2\}\}$$

Then the local related function is formed to compute all local reducts:

$$f(TS, CF, D) = \wedge(\mathcal{C}_1 \vee \mathcal{C}_2 \vee \mathcal{C}_3 \vee \mathcal{C}_4 \vee) \wedge (\mathcal{C}_1 \vee \mathcal{C}_2) \wedge (\mathcal{C}_1 \vee \mathcal{C}_3 \vee \mathcal{C}_4) \wedge (\mathcal{C}_2) = (\mathcal{C}_1 \wedge \mathcal{C}_2) \vee (\mathcal{C}_2 \wedge \mathcal{C}_3) \vee (\mathcal{C}_2 \wedge \mathcal{C}_4).$$

Thus, $RED(CF) = \{\{\mathcal{C}_1, \mathcal{C}_2\}, \{\mathcal{C}_2, \mathcal{C}_3\}, \{\mathcal{C}_2, \mathcal{C}_4\}\}$ and $CORE(CF) = \{\mathcal{C}_2\}$.

It shows that all local reducts of a covering decision system can be computed by the local related family method.

4 Local Feature Selection Algorithms

Although all local reducts can be computed by the local related function, it is proved to be NP hard to obtain all local reducts or an optimal reduct. To quickly obtain a local reduct, a heuristic algorithm based on local related family (LRF) is designed. The flowchart of LRF is shown in Figure 4.1.

LRF is divided into two steps:

Step 1: Calculate the local related family by Definite 6. For any sample w_i in the data table, the information granule of w_i induced by a condition feature is define as

$$\delta_a(w_i) = \{w_j \,||\, a(w_i) - a(w_j) | \le \delta\},$$

where $a(w_i)$ is the value of sample w_i under Feature a, $\delta \in [0,1]$ is the neighborhood radius that controls the size of all information granules.

Step 2: Calculate local reducts by greedy strategy on the basis of local related family.

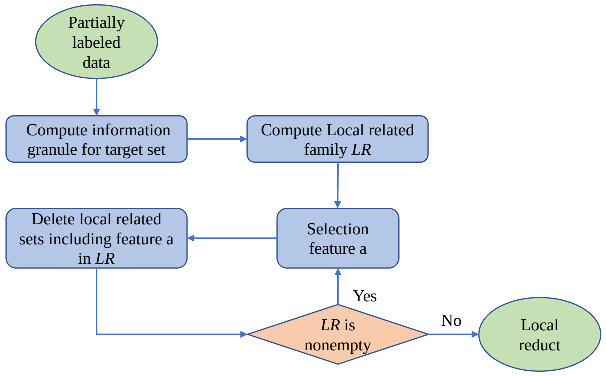

Partially labeled data
Compute information granule for target set
Compute Local related family *LR*
Selection feature a
Delete local related sets including feature a in *LR*
LR is nonempty
Yes
No
Local reduct

Fig. 4.1. The flowchart of LRF.

Algorithm 1: Local feature selection algorithm based on local related family (LRF).

Input: A data table (U, CF, D), the target set $TS \subseteq U$, a variable precision α and a neighborhood radius δ;
Output: A local reduct RED.

```
    // Step 1: Calculating the local related family.
1: For any $w_i \in TS$, let $\gamma(w_i) = \phi$ and $LR(TS, CF, D) = \phi$.
2: for $1 \le i \le m$, $a_i \in CF$, $CF = \{a_1, a_2, ..., a_m\}$
3: for $1 \le j \le n$, $w_j \in TS$, $TS = \{w_1, w_2, ..., w_n\}$
4: If $a_i \notin \gamma(w_j)$ go to 5, otherwise skip to the next cycle;
5: computing information granule $\delta_{a_i}(w_j)$;
6: if $\theta(W / \delta_{a_i}(w_j)) \ge \alpha$, where $W = [w_j]_D \bigcap TS$
7: If $w_t \in \delta_{a_i}(w_j)$ for any $w_t \in W$, then $\gamma(w_t) = \gamma(w_t) \bigcup \{a_i\}$;
8: end if
9: end for
10: end for
11: $LR(TS, CF, D) = \{\gamma(w_i) | w_i \in TS$ and $\gamma(w_i) \ne \phi\}$;
    // Step 2: Calculating local reduct by greedy strategy on the basis of local related family.
12: $RED = \phi$;
13: while $LR(TS, CF, D) \ne \phi$
14: if $a \in \bigcup LR(TS, CF, D)$ and $\|a\| = \max\{\|a\| : a \in \bigcup LR(TS, CF, D)\}$ % $\|a\|$ is the number of occurrences of $a$ in $LR(TS, CF, D)$ %
15: let $RED = RED \bigcup \{a\}$;
16: If $a \in \gamma(w_i)$ for any $\gamma(w_i) \in LR(TS, CF, D)$,
      then let $LR(TS, CF, D) = LR(TS, CF, D) - \{\gamma(w_i)\}$;
17: end if
18: end while
19: Output the local reduct $RED$, end.
```

We analyze the time and space complexities of the LRF algorithm, and compare them with existing three elegant algorithms: Local Attribute Reduction of target Decision (LARD) [49], Granular Ball Neighborhood Rough Sets (GBNRS) [33] and Global Related Family (GRF) [54].

Given a data table (U, CF, D), $TS(TS \subseteq U)$ is the set of all samples with labels. The time and space complexities of four algorithms is shown in Table 4.1.

The time complexity of LRF in Step 1 is $O(|CF||TS||U|)$, and in Step 2 is $O(\min\{|CF|, |TS|\})$. Therefore, the time complexity of LRF is $O(|CF||TS||U| + \min\{|CF|, |TS|\})$. The space complexity is $O(|CF||TS|)$.

The time complexity of the global related family algorithm GRF [54] is $O(|CF||U|^2 + \min\{|CF|, |U|\})$. And the space complexity of GRF is $O(|CF||U|)$. It is evident that the time and space complexities of LRF is lower than GRF.

The time complexity of LARD [49] is

$$O(\sum_{i=1}^{|CF|}(|CF| - i + 1)(\sum_{j=1}^{r}|W_i^{\,j}||U_i| + \sum_{j=1}^{r}|W_i^{\,j}|^2)),$$

where $U_i \subseteq U(U_1 = U)$, $TS_i \subseteq U_i(TS_1 = TS)$, $W_i^{\,j} \in TS_i / D$. And

TABLE 4.1
The Time and Space Complexities of Four Algorithms

Algorithm	Time complexities	Space complexities
LARD	$\mathrm{O}(\sum_{i=1}^{\lvert CF\rvert}(\lvert CF\rvert-i+1)(\sum_{j=1}^{r}\lvert W_i^{\,j}\rvert\lvert U_i\rvert+\sum_{j=1}^{r}\lvert W_i^{\,j}\rvert^2))$	$\mathrm{O}(\lvert CF\rvert\lvert TS\rvert)$
GBNRS	$\mathrm{O}(L\lvert CF\rvert^2\lvert U\rvert)$	$\mathrm{O}(\lvert U\rvert^2/2)$
GRF	$\mathrm{O}(\lvert CF\rvert\lvert U\rvert^2+\min\{\lvert CF\rvert,\lvert U\rvert\})$	$\mathrm{O}(\lvert CF\rvert\lvert U\rvert)$
LRF	$\mathrm{O}(\lvert CF\rvert\lvert TS\rvert\lvert U\rvert+\min\{\lvert CF\rvert,\lvert TS\rvert\})$	$\mathrm{O}(\lvert CF\rvert\lvert TS\rvert)$

the space complexity of LARD is $\mathrm{O}(|CF||TS|)$. Since

$$|CF|<\sum_{i=1}^{|CF|}(|CF|-i+1)\,,\ \sum_{j=1}^{r}|W_1^{\,j}|=|TS_1|\ \text{ and}\ |TS_1|\le\sum_{j=1}^{r}|W_1^{\,j}|^2\,,$$

the time complexity of LRF is lower than LARD.

To show the difference visually, the computation processes of LRF and LARD are compared in Fig. 1.1. During the computation of LARD, the data needs to be granulated multiple times based on multiple features, while each feature only needs to be granulated once for LRF, which significantly reduces the computation time.

The time complexity of GBNRS [33] is $\mathrm{O}(L|CF|^2|U|)$, where L is an unpredictable number，it is usually a large number, and its space complexity is $\mathrm{O}(|U|^2/2)$. The time and space complexities of LRF is lower than GBNRS. In next section, efficiency advantage of the new algorithm is verified by several experiments.

5 Experiment Analysis

In this section, numerical experiments are conducted based on 13 public datasets (downloaded from the UCI database http://archive.ics.uci.edu/ml/datasets.php and the KEEL database http://sci2s.ugr.es/keel/datasets.php) to test effectiveness and efficiency of LRF. Normalization was conducted for each data set. The description of all datasets is listed in Table 5.1 and Table 5.4.

LRF is compared with three existing algorithms: Local Attribute Reduction of target Decision (LARD) [49], Granular Ball Neighborhood Rough Sets (GBNRS) [33] and Global Related Family (GRF) [54]. It is notable that the results provided by GBNRS and GRF are computed based on raw data sets without deleting any label. In another word, LRF is compared with one local and two global algorithms. Actually, comparing LRF with global algorithms is much more challenging and authentic than comparing with some semi-supervised algorithms. (1) For classification accuracy, considering some of semi-supervised algorithms need adding pseudo labels to samples without any label, the results based on pseudo labels are usually less reliable and less accurate than those based on real labels. (2) For computation efficiency, applying global algorithms avoids generating pseudo labels, which saves a lot of time for comparing algorithms. Furthermore, LARD [49], GBNRS [33] and GRF [54] are excellent feature selection algorithms highly rated by scholars [51, 52]. Thus, the results in this paper are credible.

All experiments are implemented by MATLAB R2017a, on a PC with Windows 10 system, Intel Core i5-1035G7 CPU 1.5GHz and 8.0GB memory.

TABLE 5.1
The Information of Datasets

	Data	Samples	Features	Class
1	hill-valley	606	100	2
2	page blocks	5472	10	5
3	satimage	6435	36	7
4	anuran calls	7195	22	4
5	thyroid	7200	21	3
6	crowdsourced	10545	28	6
7	magic	19020	10	2
8	shuttle	57999	9	5
9	census	142521	41	2

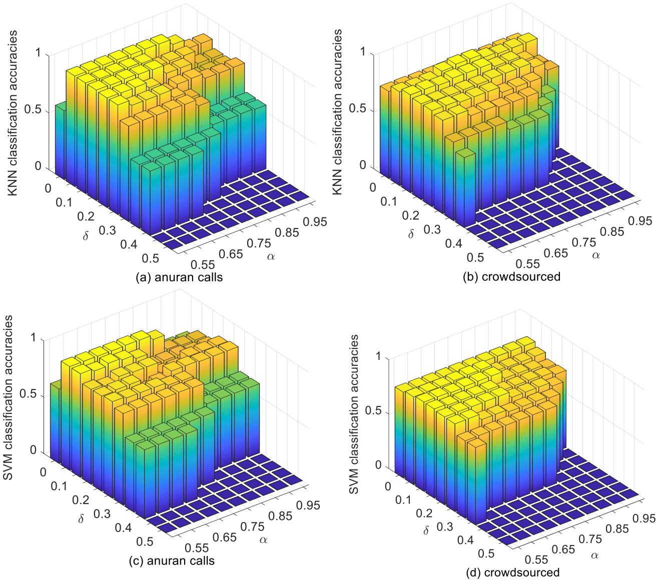


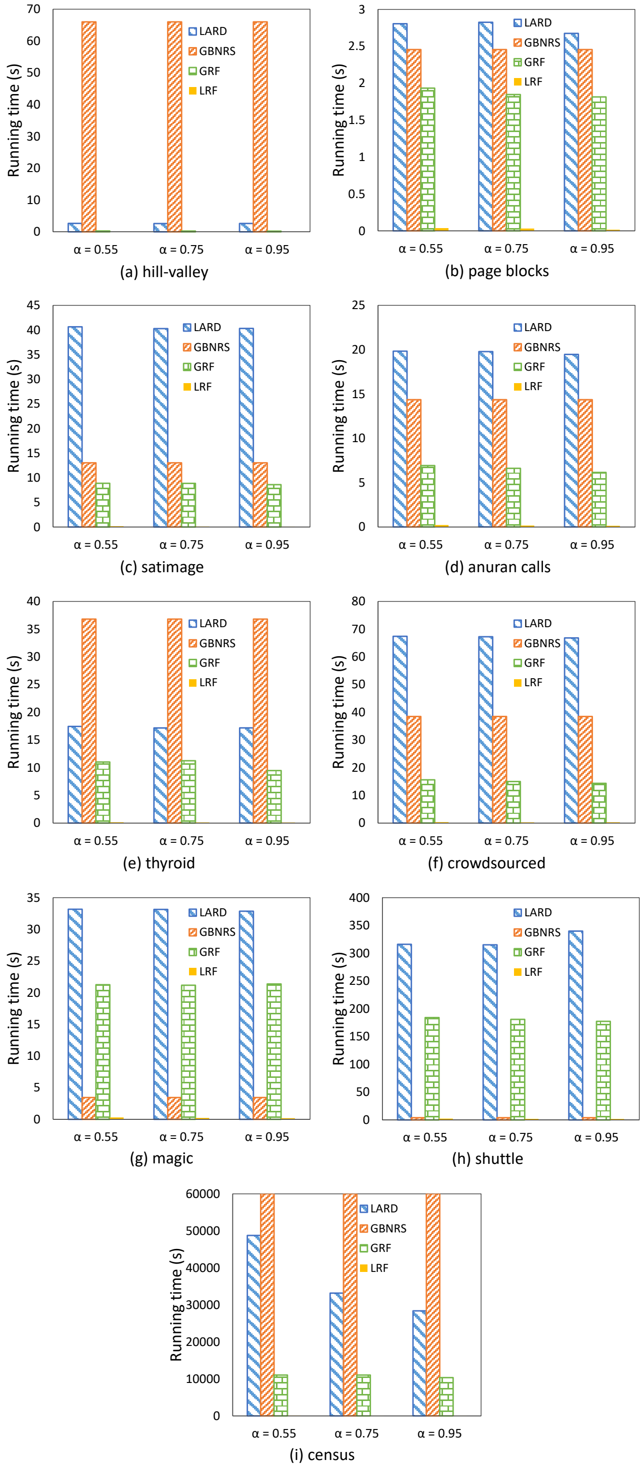
Fig. 5.1. Parameter Comparison.

5.1 Parameter analysis

In this subsection, experiments are conducted based on data sets in Table 5.1 to test different parameters of LRF. We obtain partially labeled data sets by retaining labels of the first 10% samples for each data set.

LRF has two parameters, namely, the variable precision and neighborhood radius, where the range of variable precision is $\alpha\in(0.5,\,1]$ and neighborhood radius, $\delta\in[0,\,0.5]$. Within the above range, 10 different variable precision values: 0.55, 0.60, 0.65, …, 1, and 11 different radius values: 0, 0.05, 0.1, …, 0.5, are chosen respectively. KNN(K=3) and SVM classifiers are used to evaluate the classification accuracy of all feature subsets selected by the LRF under different parameters. The experimental results are as shown in Figure 5.1.

Since results of different data sets are in similar pattern, we only show results of two data sets in Figure 5.1. As can be seen from Figure 5.1, the optimal KNN classification accuracy usually occurs when $\delta\in[0,\,0.3]$. Therefore, [0, 0.3] is the recommended value range of neighborhood radius δ . Since δ is set to 0.001 for LARD in [49], the



(a) hill-valley

(b) page blocks

(c) satimage

(d) anuran calls

(e) thyroid

(f) crowdsourced

(g) magic

(h) shuttle

(i) census

Fig. 5.2. Running time comparison of four algorithms.

radius value is also set to 0.001 in the following experiment for the sake of comparison.

For the variable precision parameters, the optimal value is between (0.5, 0.85]. Therefore, (0.5, 0.85] is the recommended value range of the variable precision for LRF. Since the better range of variable precision parameters for GRF is [0.7, 1], in the following experiments, three values (0.55, 0.75, and 0.95) are set for both LRF and GRF.

TABLE 5.2
Running Time of Four Algorithms (in seconds)

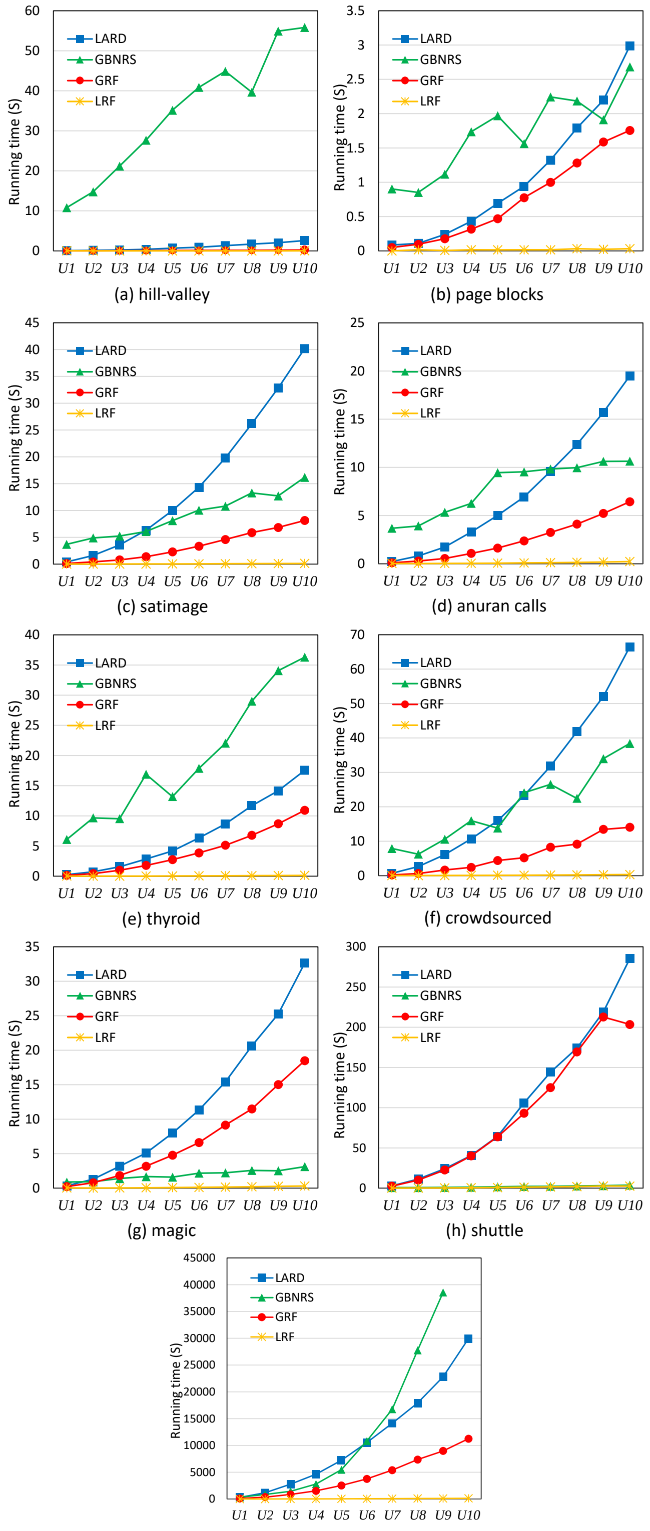
Data	α	LARD	GBNRS	GRF	LRF
1	0.55	2.650	65.995	0.263	0.016
	0.75	2.605		0.234	0.015
	0.95	2.627		0.212	0.016
2	0.55	2.806	2.456	1.933	0.037
	0.75	2.825		1.847	0.032
	0.95	2.675		1.814	0.015
3	0.55	40.648	13.029	8.878	0.153
	0.75	40.282		8.854	0.092
	0.95	40.318		8.605	0.083
4	0.55	19.834	14.353	6.919	0.218
	0.75	19.78		6.616	0.157
	0.95	19.459		6.147	0.135
5	0.55	17.435	36.812	11.01	0.138
	0.75	17.164		11.247	0.098
	0.95	17.18		9.466	0.082
6	0.55	67.397	38.479	15.611	0.338
	0.75	67.245		14.976	0.226
	0.95	66.836		14.341	0.255
7	0.55	33.171	3.434	21.247	0.297
	0.75	33.148		21.17	0.215
	0.95	32.865		21.371	0.176
8	0.55	316.12	4.002	184.278	2.258
	0.75	315.213		181.027	1.815
	0.95	339.91		177.35	1.634
9	0.55	48751.04	*	11031.64	119.919
	0.75	33164.68		11029.34	94.41
	0.95	28374.14		10330.4	85.402

5.2 Computation efficiencies

In this subsection, the running time of four algorithms is compared. To obtain partially labeled data sets, the first 10% samples of each data set in Table 5.1 were collected as the target set, and the rest samples are regarded as unlabeled.

In the experiment, the parameters are respectively set as $\alpha = 0.55$, 0.75, 0.95, and $\delta = 0.001$, with results shown in Figure 5.2 and Table 5.2. In this paper, "*" in all tables refers to insufficient memory, "\" indicates that the running time is more than 72 hours, and no result is returned in either case.

As seen from Figure 5.2 and Table 5.2, LRF is much faster than three other algorithms for all datasets. Therefore, the columns of LRF in Figure 5.2 are too short to spot. In addition, the larger the data set scale, the greater the efficiency advantage of LRF. For example, Table 5.2 shows that on Data 2 (page blocks with 5472 samples, 10 features) the running time of LRF is about 1/75 of that of LARD, on Data 5 (thyroid with 7200 samples, 21 features), 1/126, and on Data 3 (satimage with 6435 samples, 36 features), 1/265. Furthermore, on Data 9 (census with 142521 samples, 41 features), the running time of LRF is 1/406 of that of LARD and 1/92 of that of GRF, while GBNRS runs out of memory. That means LRF can run 405 times faster than LARD and 91 times faster than GRF.



(a) hill-valley

(b) page blocks

(c) satimage

(d) anuran calls

(e) thyroid

(f) crowdsourced

(g) magic

(h) shuttle

(i) census

Fig. 5.3. Comparison running time of four algorithms.

To show the different patterns of efficiency changes of four algorithms when the number of samples increases gradually, each data set in Table 1 is equally divided into ten sub-data sets. A sub-data set is randomly selected as the initial set $(U1)$ to be computed. Then one sub-data set is added to the set at each step, for example, $|U1|=|U|/10$, $|U2|=2|U|/10$, …, $|U10|=10|U|/10$. The target set is the first 10% samples of the set. At each step, the reduct of the set is computed and the running time is recorded.

The running time is shown in Figure 5.3. Compared with the other three algorithms, the computational efficiency of LRF goes up with the increase of the sample scale. On a larger data set, the computational time differences are up to hundreds of times. It is worth noting that the trend of time growth is linear, and the growth rate is very low. Meanwhile, it is obvious that, even the global algorithm GRF is much faster than LARD and GBNRS on seven data sets. Therefore, the local strategy based on related family significantly improves the computational efficiency of feature selection.

5.3 Classification accuracy

In this subsection, the effectiveness of the LRF algorithm is verified by comparing the classification accuracy of the other three algorithms. *KNN* (*K*=3) and SVM classifiers are used to evaluate the classification accuracy of different feature subsets selected by the four algorithms. As shown in Table 5.3, each result is obtained using the 10-fold cross-validation method, and the number of features selected is shown in parentheses.

Table 5.3 shows that among the 9 data sets computed by LRF, the *KNN* (*K*=3) and SVM classification accuracy of 4 data sets is the highest. The average *KNN* (*K*=3) and SVM classification accuracy of LRF is basically the same as global algorithms GBNRS and GRF, and is slightly higher than the local algorithm LARD. Evidently, LRF is an effective feature selection method for partially labeled data sets.

5.4 Testing for large-scale data sets

To compare the performances of four algorithms on large-scale data sets, two data sets, SUSY and Gene expression cancer (GEC) are downloaded from UCI Database. Where SUSY is a large sample data set with 5000000 samples, 18 features and 2 classes of labels; GEC is an ultrahigh dimension data set with 801 samples, 20531 features and 5 classes of labels. By extracting 4000000 and 3000000 samples from SUSY, we get SUSY1 and SUSY2. The description of the data sets is shown in Table 5.4.

In this experiment, the first 10% samples of data sets SUSY, SUSY1 and SUSY2 are set as the target sets. Whereas for GEC, since the sample scale of it is relatively small, the first 50% samples are set as the target set.

Considering that the scale of SUSY, SUSY1 and SUSY2 is very large, reasonable parameters are selected for each algorithm based on the results in Subsection 5.2, so that each algorithm is run once only on the three data sets. To get better performance, the variable precision parameter for Algorithms LRF and LARD is set as $\alpha = 0.55$, and for Algorithm GRF in this subsection, as $\alpha = 0.95$. The neighborhood parameter is set as $\delta = 0.001$.

Since the sample scale of GEC is small, the values of radius δ from 0.001 to 0.05 with a step size of 0.001 are tested. Altogether 50 results are recorded for GRF and LRF, and the running time in Table 5.5 is the total time of 50 runs. Since LARD runs for more than 72 hours at a time, no LARD results are recorded. And as GBNRS has no radius parameter and can only run twice in 72 hours,

TABLE 5.3
Classification Accuracy of Reduct Data

Data	α	*K*NN				SVM			
		LARD	GBNRS	GRF	LRF	LARD	GBNRS	GRF	LRF
1	0.55	0.5382(2)	0.5297(13)	0.5329(5)	**0.5448**(2)	**0.5116**(2)	0.5033(13)	0.5049(5)	0.5016(2)
	0.75	0.5382(2)		**0.5412**(6)	0.5330(3)	**0.5116**(2)		0.5000(6)	0.5081(3)
	0.95	**0.5382**(2)		0.5282(11)	0.5330(3)	**0.5116**(2)		0.5034(11)	0.5081(3)
2	0.55	0.9912(3)	**0.9929**(6)	0.9629(1)	0.9921(7)	0.9479(3)	0.9655(6)	0.8708(1)	**0.9741**(7)
	0.75	0.9912(3)		0.9894(2)	0.9923(6)	0.9479(3)		0.9069(2)	**0.9739**(6)
	0.95	0.9912(3)		0.9921(4)	0.9923(6)	0.9479(3)		0.9611(4)	**0.9739(**6)
3	0.55	0.9857(6)	0.9880(16)	0.9800(3)	**0.9905**(19)	0.9467(6)	0.9577(16)	0.9229(3)	0.9549(19)
	0.75	0.9857(6)		0.9887(6)	**0.9902**(23)	0.9467(6)		**0.9772**(6)	0.9624(23)
	0.95	0.9857(6)		**0.9905**(24)	0.9901(16)	0.9467(6)		0.9571(24)	**0.9632**(16)
4	0.55	0.8460(4)	0.9857(14)	0.8767(2)	0.9830(9)	0.6646(4)	0.9197**(**14)	0.8389(2)	0.8866(9)
	0.75	0.8460(4)		0.9736(8)	0.9833(11)	0.6646(4)		0.8693(8)	0.8746(11)
	0.95	0.8460(4)		**0.9914**(22)	0.9882(16)	0.6646(4)		**0.9338**(22)	0.9077(16)
5	0.55	0.9255(8)	0.9393(16)	0.9258(1)	0.9264(3)	**0.9258**(8)	**0.9258**(16)	**0.9258**(1)	**0.9258**(3)
	0.75	0.9255(8)		0.9258(1)	0.9264(3)	**0.9258**(8)		**0.9258**(1)	**0.9258**(3)
	0.95	0.9255(8)		**0.9469**(10)	0.9264(3)	0.9258(8)		**0.9271**(10)	0.9258(3)
6	0.55	0.7921(3)	0.9687(26)	0.8720(2)	0.9460(15)	0.7991(3)	0.8612(26)	0.8015(2)	0.8418(15)
	0.75	0.7921(3)		0.9393(12)	0.9406(14)	0.7991(3)		0.8277(12)	0.8601(14)
	0.95	0.7921(3)		**0.9696**(28)	0.9431(14)	0.7991(3)		**0.8739**(28)	0.8521(14)
7	0.55	**0.9984**(2)	**0.9984**(3)	**0.9984**(2)	**0.9984**(3)	**0.9982**(2)	**0.9982**(3)	**0.9982**(2)	**0.9982**(3)
	0.75	**0.9984**(2)		**0.9984**(2)	**0.9984**(3)	**0.9982**(2)		**0.9982**(2)	**0.9982**(3)
	0.95	**0.9984**(2)		**0.9984**(2)	**0.9984**(3)	**0.9982**(2)		**0.9982**(2)	**0.9982**(3)
8	0.55	0.9997(6)	**0.9998**(5)	0.9355(1)	**0.9998**(8)	0.9598(6)	**0.9698**(5)	0.8696(1)	0.9598(8)
	0.75	0.9997(6)		0.9707(2)	**0.9998**(8)	0.9598(6)		0.8697(2)	0.9598(8)
	0.95	0.9997(6)		**0.9998**(8)	**0.9998**(8)	0.9598(6)		0.9598(8)	0.9598(8)
9	0.55	**0.9412**(19)	*	0.9378(1)	0.9309(3)	**0.9478**(19)	*	0.9427(1)	0.9429(3)
	0.75	**0.9412**(19)		0.9378(1)	0.9309(3)	**0.9478**(19)		0.9427(1)	0.9429(3)
	0.95	0.9412(19)		**0.9417**(17)	0.9309(3)	0**.9478**(19)		0.9458(17)	0.9429(3)
average		0.8909(5.89)	**0.9253**(12.37)	0.9128(6.81)	0.9226(7.96)	0.8557(5.89)	0.8876(12.37)	0.8723(6.81)	**0.8898**(7.96)

two results are recorded, and the running time in Table 5.5 is the total time of two runs. Then, the results with the highest sum of classification accuracy evaluated by *K*NN and SVM classifiers for each algorithm are selected. The results are shown in Table 5.5 and Table 5.6, with the numbers of features in parentheses.

Table 5.5 shows the running time of four algorithms. For Data GEC, even if GBNRS only runs twice, it is 173.6 times of the total time of LRF running 50 times. The running time of GRF is 3.89 times that of LRF. LARD algorithm cannot complete the computing of initial parameters within 72 hours, so it does not return any results. In the computing of large sample data SUSY, SUSY1 and SUSY2, LARD and GRF fail to return any results because the running time exceeds 72 hours. GBNRS cannot compute due to insufficient memory and does not return any results, and only LRF can successfully compute all feature reducts. As can be seen from Table 5.6, the classification accuracy of the data after LRF reduction is basically the same as that of the original data, or it is significantly improved.

Next, the actual scale of the data that can be computed by the four algorithms on the personal computer is tested and compared on the data set SUSY. Under the fixed 18 feature dimensions, the sample size is gradually increased, and the computing time is limited to 8-10 hours. Among them, LRF takes 31008 seconds (about 8.6 hours) to compute the data set of five million samples; GRF takes 33,741 seconds (about 9.4 hours) to compute the data set of 500,000 samples; LARD takes 32009 seconds (about 8.9 hours) to compute the data set of 350,000 samples. Although it takes 9040 seconds for GBNRS to compute 100,000 sample data, when the sample reaches 110,000, GBNRS cannot continue to compute due to insufficient memory. The comparison results of the computing power of the four algorithms are shown in Figure 5.4. As can be seen, the computing power of LRF far exceeds the other three algorithms. Therefore, LRF algorithm can quickly and efficiently process partially labeled large-scale data with limited resources.

TABLE 5.4
The Information of Datasets

Data	Samples	Features	Class
Gene expression cancer (GEC)	801	20531	5
SUSY	5000000	18	2
SUSY1	4000000	18	2
SUSY2	3000000	18	2

TABLE 5.5
Running Time of Four Algorithms

Algorithm		GEC 801*20532	SUSY2 3000000*19	SUSY1 4000000*19	SUSY 5000000*19
LARD	δ	0.001	0.001	0.001	0.001
	time	\	\	\	\
GBNRS	time	237258.73	*	*	*
GRF	δ	0.005	0.001	0.001	0.001
	time	5313.39	\	\	\
LRF	δ	0.01	0.001	0.001	0.001
	time	1366.46	9944.62	17191.24	31008.28

TABLE 5.6
Classification Accuracy of Reduct Data

Data	Classifier	Original	LARD	GBNRS	GRF	LRF
GEC 801*20532	*K*NN	0.9947(20531)	\	0.8801(29)	0.9988(10)	0.9975(12)
	SVM	0.9912(20531)	\	0.9200(29)	0.9988(10)	0.9975(12)
SUSY2 3000000*19	*K*NN	0.6447(18)	\	*	\	0.7463(8)
	SVM	0.6998(18)	\	*	\	0.7615(8)
SUSY1 4000000*19	*K*NN	0.6444(18)	\	*	\	0.7450(8)
	SVM	0.7859(18)	\	*	\	0.7624(8)
SUSY 5000000*19	*K*NN	0.7299(18)	\	*	\	0.7496(9)
	SVM	0.7928(18)	\	*	\	0.7712(9)

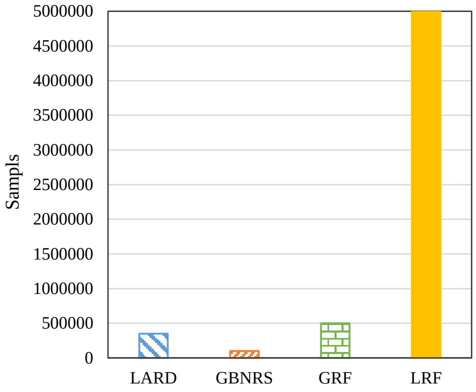


Fig. 5.4. Comparison of computing capacities.

6 Conclusion And Future Work

In order to improve the computational efficiency of the feature selection and process partially labeled large-scale data, a new feature evaluation method, namely, the local related family, is proposed in this paper. This method can evaluate features more efficiently by computing only the relevant information of labeled samples instead of computing all samples. Then a local feature selection algorithm LRF with linear time and space complexities is designed. The experiment results show that, compared with the other three effective algorithms, this algorithm improves the computational efficiency by even 405 times and achieves good performance in processing partially labeled large-scale data sets. However, there are limitations in processing small sample partially labeled data sets. Therefore, how to make full use of the potential labeling information of unlabeled samples and how to improve the classification accuracy based on semi-supervised learning method need to be further studied.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No.11201490, No.61976089 and No.61673388), the Natural Science Foundation of Hunan Province under Grant (No.2021JJ20037), the Training Program for Excellent Young Innovators of Changsha (No. kq1905031), the Key R&D Program of Hunan Province (No.2018SK2129), and the Hunan Provincial Science & Technology Project Foundation (2018TP1018).

References

[1] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in Proceedings of the Eleventh Annual Conference on Computational Learning Theory, 1998, pp. 92-100.

[2] T. Joachims, "Transductive inference for text classification using support vector machines," in Proceedings of the Sixteenth International Conference on Machine Learning, 1999, pp. 200-209.

[3] X. Zhu, Z. Ghahramani, and J. Lafferty, "Learning from Labeled and Unlabeled Data with Label Propagation," Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

[4] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in Proceedings of the Twentieth International conference on Machine learning, 2003, pp. 912-919.

[5] C. Hou and Z. Zhou, "One-Pass Learning with Incremental and Decremental Features," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 11, pp. 2776-2792, 2018.

[6] C. Hou, L. Zeng, and D. Hu, "Safe Classification with Augmented Features," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 9, pp. 2176-2192, 2019.

[7] M. Wang, W. Fu, X. He, S. Hao, and X. Wu, "A Survey on Large-scale Machine Learning," IEEE Transactions on Knowledge and Data Engineering, 2021.

[8] M. Wang, W. Fu, S. Hao, D. Tao, and X. Wu, "Scalable Semi-Supervised Learning by Efficient Anchor Graph Regularization," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 7, pp. 1864-1877, 2016.

[9] M. Wang, W. Fu, S. Hao, H. Liu, and X. Wu, "Learning on Big Graph: Label Inference and Regularization with Anchor Hierarchy," IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 5, pp. 1101-1114, 2017.

[10] Z. Yu et al., "Progressive Semisupervised Learning of Multiple Classifiers," IEEE Transactions on Cybernetics, vol. 48, no. 2, pp. 689-702, 2018.

[11] Z. Yu et al., "Multiobjective Semisupervised Classifier Ensemble," IEEE Transactions on Cybernetics, vol. 49, no. 6, pp. 2280-2293, 2019.

[12] Z. Yu et al., "Adaptive Semi-Supervised Classifier Ensemble for High Dimensional Data Classification," IEEE Transactions on Cybernetics, vol. 49, no. 2, pp. 366-379, 2019.

[13] B. Xue, M. Zhang, and W. N. Browne, "Particle Swarm Optimization for Feature Selection in Classification: a Multi-Objective Approach," IEEE Transactions on Cybernetics, vol. 43, no. 6, pp. 1656-1671, 2013.

[14] X. Zhang, G. Wu, Z. Dong, and C. Crawford, "Embedded feature-selection support vector machine for driving pattern recognition," Journal of the Franklin Institute, vol. 352, no. 2, pp. 669-685, 2015.

[15] P. Bermejo, J. A. Gamez, and J. M. Puerta, "Speeding up incremental wrapper feature subset selection with Naive Bayes classifier," Knowledge-Based Systems, vol. 55, pp. 140-147, 2014.

[16] Z. Deng, F. Chung, and S. Wang, "Robust Relief-Feature Weighting, Margin Maximization, and Fuzzy Optimization," IEEE Transactions on Fuzzy Systems, vol. 18, no. 4, pp. 726-744, 2010.

[17] X. Zhu, S. Zhang, R. Hu, and Y. Zhu, "Local and Global Structure Preservation for Robust Unsupervised Spectral Feature Selection," IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 3, pp. 517-529, 2017.

[18] J. Xiao, H. Cao, X. Jiang, X. Gu, and L. Xie, "GMDH-based semi-supervised feature selection for customer classification," Knowledge-Based Systems, vol. 132, pp. 236-246, 2017.

[19] H. Zhang, J. Wang, Z. Sun, J. M. Zurada, and N. R. Pal, "Feature Selection for Neural Networks Using Group Lasso Regularization," IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 4, pp. 659-673, 2019.

[20] S. Maldonado and R. Weber, "A wrapper method for feature selection using support vector machines," Information Sciences, vol. 179, no. 13, pp. 2208-2217, 2009.

[21] W. Pedrycz, "Granular Computing for Data Analytics: A Manifesto of Human-Centric Computing," IEEE/CAA Journal of Automatica Sinica, vol. 5, no. 6, pp. 1025-1034, 2018.

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

[23] Z. Pawlak, "Rough Sets: Theoretical Aspects of Reasoning about Data," Kluwer Academic Publishers, 1992.

[24] T. Yang, X. Zhong, G. Lang, Y. Qian, and J. Dai, "Granular Matrix: A New Approach for Granular Structure Reduction and Redundancy Evaluation," IEEE Transactions on Fuzzy Systems, vol. 28, no. 12, pp. 3133-3144, 2020.

[25] Q. Hu, L. Zhang, Y. Zhou, and W. Pedrycz, "Large-Scale Multi-Modality Attribute Reduction with Multi-Kernel Fuzzy Rough Sets," IEEE Transactions on Fuzzy Systems, vol. 26, no. 1, pp. 226-238, 2018.

[26] J. Chen, J. Mi, and Y. Lin, "A graph approach for fuzzy-rough feature selection," Fuzzy Sets and Systems, vol. 391, pp. 96-116, 2020.

[27] J. Hu, T. Li, H. Wang, and H. Fujita, "Hierarchical cluster ensemble model based on knowledge granulation," Knowledge-Based Systems, vol. 91, pp. 179-188, 2016.

[28] J. Mi, W. Wu, and W. Zhang, "Approaches to knowledge reduction based on variable precision rough set model," Information Sciences, vol. 159, no. 3-4, pp. 255-272, 2004.

[29] J. Zhang, J. Wong, Y. Pan, and T. Li, "A Parallel Matrix-Based Method for Computing Approximations in Incomplete Information Systems," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 2, pp. 326-339, 2015.

[30] J. Chen, Y. Lin, G. Lin, J. Li, and Y. Zhang, "Attribute reduction of covering decision systems by hypergraph model," Knowledge-Based Systems, vol. 118, pp. 93-104, 2017.

[31] C. Wang, Q. Hu, X. Wang, D. Chen, Y. Qian, and Z. Dong, "Feature Selection Based on Neighborhood Discrimination Index," IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 7, pp. 2986-2999, 2018.

[32] L. Sun, X. Zhang, Y. Qian, J. Xu, and S. Zhang, "Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification," Information Sciences, vol. 502, pp. 18-41, 2019.

[33] S. Xia, Z. Zhang, W. Li, G. Wang, E. Giem, and Z. Chen, "GBNRS: A Novel Rough Set Algorithm for Fast Adaptive Attribute Reduction in Classification," IEEE Transactions on Knowledge and Data Engineering, pp. 1041-4347, 2020.

[34] D. Chen, S. Zhao, L. Zhang, Y. Yang, and X. Zhang, "Sample Pair Selection for Attribute Reduction with Rough Set," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 11, pp. 2080-2093, 2012.

[35] H. Chen, T. Li, X. Fan, and C. Luo, "Feature selection for imbalanced data based on neighborhood rough sets," Information Sciences, vol. 483, pp. 1-20, 2019.

[36] J. Liang, F. Wang, C. Dang, and Y. Qian, "A Group Incremental Approach to Feature Selection Applying Rough Set Technique," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 2, pp. 294-308, 2014.

[37] P. Maji, "A Rough Hypercuboid Approach for Feature Selection in Approximation Spaces," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 16-29, 2014.

[38] H. Chen, T. Li, Y. Cai, C. Luo, and H. Fujita, "Parallel attribute reduction in dominance-based neighborhood rough set," Information Sciences, vol. 373, pp. 351-368, 2016.

[39] J. Dai and J. Chen, "Feature selection via normative fuzzy information weight with application into tumor classification," Applied Soft Computing, vol. 92, p. 106299, 2020.

[40] J. Dai, J. Chen, Y. Liu, and H. Hu, "Novel multi-label feature selection via label symmetric uncertainty correlation learning and feature redundancy evaluation," Knowledge-Based Systems, vol. 207, p. 106342, 2020.

[41] J. Zhang, G. Zhang, Z. Li, L. Qu, and C.-F. Wen, "Feature selec-

tion in a neighborhood decision information system with application to single cell RNA data classification," Applied Soft Computing, vol. 113, p. 107876, 2021.

[42] S. Luo, D. Miao, Z. Zhang, Y. Zhang, and S. Hu, "A neighborhood rough set model with nominal metric embedding," Information Sciences, vol. 520, pp. 373-388, 2020.

[43] Q. Hu, D. Yu, J. Liu, and C. Wu, "Neighborhood rough set based heterogeneous feature subset selection," Information Sciences, vol. 178, no. 18, pp. 3577-3594, 2008.

[44] 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.

[45] J. Dai and Q. Xu, "Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification," Applied Soft Computing, vol. 13, no. 1, pp. 211-221, 2013.

[46] J. Dai, Q. Hu, J. Zhang, H. Hu, and N. Zheng, "Attribute selection for partially labeled categorical data by rough set approach," IEEE Transactions on Cybernetics, vol. 47, no. 9, pp. 2460-2471, 2017.

[47] K. Liu, X. Yang, H. Yu, J. Mi, P. Wang, and X. Chen, "Rough set based semi-supervised feature selection via ensemble selector," Knowledge-Based Systems, vol. 165, pp. 282-296, 2019.

[48] Y. Qian et al., "Local rough set: a solution to rough data analysis in big data," International Journal of Approximate Reasoning, vol. 97, pp. 38-63, 2018.

[49] Q. Wang, Y. Qian, X. Liang, Q. Guo, and J. Liang, "Local neighborhood rough set," Knowledge-Based Systems, vol. 153, pp. 53-64, 2018.

[50] T. Yang, Q. Li, and B. Zhou, "Related family: a new method for attribute reduction of covering information systems," Information Sciences, vol. 228, pp. 175-191, 2013.

[51] G. Lang, M. Cai, H. Fujita, and Q. Xiao, "Related families-based attribute reduction of dynamic covering decision information systems," Knowledge-Based Systems, vol. 162, pp. 161-173, 2018.

[52] G. Lang, Q. Li, M. Cai, H. Fujita, and H. Zhang, "Related families-based methods for updating reducts under dynamic object sets," Knowledge and Information Systems, vol. 60, no. 2, pp. 1081-1104, 2019.

[53] M. Cai, G. Lang, H. Fujita, Z. Li, and T. Yang, "Incremental approaches to updating reducts under dynamic covering granularity," Knowledge-Based Systems, vol. 172, pp. 130-140, 2019.

[54] B. Ou, H. Zhang, J. Dai, and T. Yang, "Intrusion detection method based on variable precision covering rough set," Journal of Computer Applications, vol. 40, no. 12, pp. 3465-3470, 2020.

[55] Z. Bonikowski, E. Bryniarski, and U. Wybraniec-Skardowska, "Extensions and intentions in the rough set theory," Information Sciences, vol. 107, no. 1-4, pp. 149-167, 1998.

[56] E. Bryniarski, "A calculus of rough sets of the first order," Bulletin of the Polish Academy of Sciences. Mathematics, vol. 37, no. 1-6, pp. 71-78, 1989.

[57] X. Zheng and J. Dai, "A variable precision covering generalized rough set model," in International Conference on Rough Sets and Knowledge Technology, 2011, pp. 120-125: Springer.

[58] Y. Zhang and Y. Wang, "Covering rough set model based on variable precision," Journal of Liaoning Institute of Technology, vol. 26, no. 4, pp. 274-276, 2006.

Tian Yang received the Ph.D. degree from Hunan University, Changsha, China, in Applied Mathematics. She is an associate professor at Hunan Normal University, Changsha, China. Her current research areas include granular computing, intelligent information processing, fuzzy Systems, Data Mining and topology.

Yanfang Deng is currently pursuing the Master Degree with the College of Information Science and Engineering, Hunan Normal University, Changsha, China. Her main research interests include granular computing and data mining.

Bin Yu received the Ph.D. degree from Hunan University, Changsha, China, in Applied Mathematics. He is a teacher at Hunan Normal University, Changsha, China. His current research areas include granular computing, intelligent information processing, fuzzy Systems, Data Mining and topology.

Yuhua Qian received the M.S. degree and the Ph.D. degree in computers with applications from Shanxi University in 2005 and 2011, respectively. He is currently a Professor with the School of Computer and Information Technology, Shanxi University, Taiyuan, China. His current research interests include pattern recognition, machine learning, rough sets, granular computing and artificial intelligence.

Jianhua Dai received the B.Sc., M.Eng., and Ph.D. degrees in computer science from Wuhan University, Wuhan, China, in 1998, 2000, and 2003, respectively. He is currently the Director of Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, and the Dean of the College of Information Science and Engineering, Hunan Normal University, Changsha, China. His current research interests include artificial intelligence, machine learning, intelligent information processing, rough sets, granular computing and neural networks.