Granular Computing (2020) 5:239–250
https://doi.org/10.1007/s41066-018-00151-5

ORIGINAL PAPER

Neighborhood attribute reduction approach to partially labeled data

Keyu Liu[1] · Eric C. C. Tsang[2] · Jingjing Song[1] · Hualong Yu[1] · Xiangjian Chen[1] · Xibei Yang[1]

Received: 23 July 2018 / Accepted: 4 December 2018 / Published online: 11 December 2018

Abstract
Presently, from the viewpoint of rough set, most of the attribute reductions are only suitable for analyzing samples with complete labels. However, in many real-world applications, it is difficult to acquire the detailed labels of all samples, it follows that many attribute reductions may be ineffective for data with both labeled and unlabeled samples, i.e., partially labeled data. To fill such a gap, the attribute reduction is explored by neighborhood rough set over partially labeled data. First, two different measurements are combined for evaluating the importance of attribute, which comes from the labeled and unlabeled samples, respectively. Second, a heuristic algorithm is re-designed using such combined importance for computing reduct. Finally, by considering several different ratios of missing labels over UCI datasets, the experimental results demonstrate that the reducts derived by our approach not only reduce the degree of uncertainty, but also offer us better classification performance. Therefore, the main contribution of this paper is to construct an effective attribute reduction strategy for partially labeled data. Moreover, this research also suggests new applications for considering attribute reduction problems in complex data.

Keywords Attribute reduction · Missing labels · Neighborhood rough set · Partially labeled data

1 Introduction

Rough set theory, introduced by Pawlak (1992), is an effective tool for handling the vagueness, imprecision and uncertainty (Zadeh 1965) in data. With more than 30 years of development, such method has been widely applied to Feature Selection (Min and Xu 2016; Swiniarski and Skowron 2003; Wang et al. 2018), Pattern Recognition (Dai et al. 2013; Hu et al. 2016), Granular Computing (Huang and Li 2018; Pedrycz and Chen 2011, 2015; Peter et al. 2003; Polkowski and Artiemjew 2015; Wang 2017; Wang et al. 2017; Zhi and Li 2018), Knowledge Discovery (Mi et al. 2004; Wu et al. 2016) and so on. Specially, as what has been pointed out by Chen et al. (2012), attribute reduction (Ju et al. 2017; Xu et al. 2016) has been considered as one of the most representative topics in rough set theory, which can be distinguished from other techniques of Feature Selection. This is mainly because: (1) attribute reductions have clear semantic explanations; (2) many measurements developed in rough set can be used to design constraints in attribute reductions.

Presently, it has been reported that attribute reduction aims to remove the redundant attributes with a given constraint. Various measurements such that approximate quality (Pawlak and Skowron 2007), conditional entropy (Hu et al. 2006) and so on have been employed to define constraints. Note that most of the measurements are generally derived from the relationship between conditional attributes and decision attribute. Therefore, the values over decision attribute, i.e., labels are required for exploring attribute reduction. From this point of view, most of the attribute reductions may only be performed over data without missing labels.

✉ Xibei Yang
jsjxy_yxb@just.edu.cn

Keyu Liu
just_liukeyu@163.com

Eric C. C. Tsang
cctsang@must.edu.mo

Jingjing Song
songjingjing108@163.com

Hualong Yu
yuhualong@just.edu.cn

Xiangjian Chen
cxj831209@163.com

[1] School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212003, Jiangsu, People's Republic of China

[2] Faculty of Information Technology, Macau University of Science and Technology, Macau 519020, People's Republic of China

In many real-world applications (Chen and Chang 2011; Chen and Chen 2009; Chen and Tanuwijaya 2011; Wang and Chen 2008), acquiring complete data is a difficult task. Generally speaking, the difficulties in labeling samples include two aspects. On the one hand, the correct labels may be unknown. For example, when adopting computer technology to aid medical institution in analyzing medical image, it is practically impossible for doctors or even experienced experts to locate the nidus (Zhou and Li 2010). Consequently, the cases of illness (labels) may be unknown. On the other hand, labeling samples is hard, expensive and time consuming as it costs too many efforts. For instance, in this era abounding with social media and connectivity, web users are becoming increasingly obsessed with interacting and sharing with their phones or computers. Along with this process, data are becoming extremely complex and it is hard for network supervisor to label all samples. To sum up, data with both labeled and unlabeled samples can be seen everywhere, and such type of data is referred to as the partially labeled data (Dai et al. 2017; Liu et al. 2018).

Up to now, most of the previous attribute reductions fail to consider the partially labeled data. Therefore, the motivations of this paper are: (1) how to handle the partially labeled data for realizing attribute reduction; (2) how to preserve or even improve the various performances of the reducts derived from the partially labeled data. For such reasons, we propose a novel attribute reduction approach to partially labeled data. The main process of our approach includes two steps: (1) considering that the approximate quality and binary relation can be used to evaluate the importance of attribute, which comes from labeled and unlabeled samples, respectively, the new importance can be expressed by combining such two measurements; (2) the significance function [also fitness function (Yang and Yao 2018)] can be constructed based on such new importance, correspondingly, a heuristic algorithm can be re-designed for computing reduct over partially labeled data.

Note that the neighborhood rough set (Hu et al. 2008b) is employed to realize the topic addressed in this paper. It is mainly because compared with other rough sets, neighborhood rough set has obvious advantages. Take the comparison between neighborhood rough set and fuzzy rough set (Dubois and Prade 1990) as an example, it is well known that (1) similar to classical rough set, the approximations obtained by neighborhood rough set are clear, while those obtained by fuzzy rough set are still fuzzy; (2) neighborhood rough set provides us a framework for handling continuous or even mixed data (Hu et al. 2008a), while fuzzy rough set is limited to continuous data.

The main contribution of our research includes two aspects: (1) a new importance for attribute reduction over partially labeled data is proposed, a heuristic algorithm based on such importance can be re-designed for computing reduct over partially labeled data; (2) through introducing the neighborhood rough set into our approach, the experimental results over several UCI datasets demonstrate that our approach is effective in selecting qualified attributes from partially labeled data.

The rest of this paper is organized as follows. Section 2 reviews some basic notations of rough set and definition of attribute reduction. The new importance for partially labeled data and the corresponding neighborhood attribute reduction approach are presented in Sect. 3. Experiments are conducted and the experimental results are analyzed in Sect. 4. Finally, we then conclude with some remarks and perspectives for future work in Sect. 5.

2 Preliminary knowledge

2.1 Neighborhood rough set

As one of the most important expanded models of classical rough set (Dou et al. 2016; Skowron and Stepaniuk 1996; Wojna 2005; Xu et al. 2017; Yang et al. 2011a, b), neighborhood rough set has been widely concerned. Since different radii used in neighborhood rough set may characterize the similarity between samples through different scales, neighborhood rough set is then more flexible and more adaptive for complex data (Hu et al. 2008a).

In rough set theory, a decision system can be described by a pair such that $\mathrm{DS} = \langle U, AT \cup \{d\}\rangle$, in which the universe U is a nonempty and finite set of samples, AT is a nonempty and finite set of conditional attributes and d is the decision attribute. Furthermore, $\forall x \in U$, $d(x)$ indicates the label of sample x.

Given a decision system DS, we assume that the values of decision attribute are discrete, then an equivalence relation over d can be defined such that $\mathrm{IND}_d = \{(x, y) \in U \times U : d(x) = d(y)\}$. By IND_d, a partition $U/\mathrm{IND}_d = \{X_1, X_2, \ldots, X_n\}$ is derived. In rough set theory, $X_k \in U/\mathrm{IND}_d$ is called the k-th decision class. Specially, the decision class which contains sample x is denoted by $[x]_d$.

Given a decision system DS, $\forall x \in U$ and $\forall A \subseteq AT$, then given a radius $\sigma \in [0, 1]$, the size of neighborhood of x related to A can be defined as (Hu et al. 2008b):

$$\sigma_A(x) = \min_{y \in U \wedge y \neq x} \triangle_A(x, y) + \sigma \cdot \left(\max_{y \in U \wedge y \neq x} \triangle_A(x, y) - \min_{y \in U \wedge y \neq x} \triangle_A(x, y) \right), \tag{1}$$

where $\triangle_A$ is one distance function, and the Euclidean distance is employed to derive $\triangle_A(x, y)$ in this paper. Note that the construction of Eq. (1) aims to avoid an undesirable case: given a sample $x \in U$, the neighborhood of x may contain only x itself if smaller radius is used. That is, any two samples can be distinguished to each other and then it is meaningless for learning process.

Following Eq. (1), the neighborhood relation in DS can be defined as:

$$\delta_A = \{(x, y) \in U \times U : \triangle_A(x, y) \leq \sigma_A(x)\}. \tag{2}$$

Correspondingly, the neighborhood of x related to A can be defined as:

$$\delta_A(x) = \{y \in U : \triangle_A(x, y) \leq \sigma_A(x)\}. \tag{3}$$

Definition 1 Given a decision system DS, $\forall A \subseteq AT$ and $\forall X_k \in U/\mathrm{IND}_d$, the neighborhood lower and upper approximations of X_k in terms of A are defined as:

$$\underline{X_k}_A = \{x \in U : \delta_A(x) \subseteq X_k\}; \tag{4}$$

$$\overline{X_k}_A = \{x \in U : \delta_A(x) \cap X_k \neq \emptyset\}. \tag{5}$$

The pair $[\underline{X_k}_A, \overline{X_k}_A]$ is called a neighborhood rough set of X_k.

2.2 Some measurements

Up to now, many measurements have been proposed to describe the certainty or uncertainty in data from the viewpoint of rough set. Similar to classical rough set, approximate quality (Pawlak 1992) can also be used for describing the degree of certainty in neighborhood rough set. The definition is as follows.

Definition 2 Given a decision system DS, $\forall A \subseteq AT$, the approximate quality of d related to A is defined as:

$$\gamma_A(d) = \frac{\left|\bigcup_{k=1}^{n} \underline{X_k}_A\right|}{|U|}, \tag{6}$$

where $|X|$ denotes the cardinality of set X.

Obviously, $0 \leq \gamma_A(d) \leq 1$ holds. If the value of approximate quality is higher, then the certainty in data is regarded as higher.

Besides approximate quality, conditional entropy is also another widely accepted measurement in rough set, which can characterize the uncertainty. Presently, many definitions of conditional entropy (Hu et al. 2010; Wei et al. 2013; Zhang et al. 2016; Zhu and Wen 2012) have been proposed with respect to different requirements. A typical representation of conditional entropy (Hu et al. 2006) is shown in Definition 3.

Definition 3 Given a decision system DS, $\forall A \subseteq AT$, the conditional entropy of d related to A is defined as:

$$\mathrm{ENT}_A(d) = -\frac{1}{|U|} \sum_{x \in U} \log \frac{|\delta_A(x) \cap [x]_d|}{|\delta_A(x)|}. \tag{7}$$

Different from approximate quality, it is not difficult to observe that the certainty is higher when the value of conditional entropy is lower.

2.3 Neighborhood classifier

Classifier can be used to evaluate the generalization performance of attributes (Chen et al. 2001). In neighborhood rough set, the neighborhood classifier proposed by Hu et al. (2008b) is frequently used. Given a test sample, neighborhood classifier uses the majority rule over labels of neighbors to determine the label of such test sample. The detailed process is shown in Algorithm 1.

Algorithm 1. Neighborhood Classifier (NEC)

Inputs: DS, test sample $y \notin U$ and radius σ;

Outputs: Predicted decision label of y : $\mathrm{Pre}_A(y)$.

1. $\forall x \in U$, compute $\triangle_A(y, x)$;
2. Obtain $\delta_A(y)$;
   `// Notes that in NEC,` $y \notin \delta_A(y)$;
3. $\forall X_k \in U/\mathrm{IND}_d$, compute the probability $\Pr(X_k|\delta_A(y)) = |\delta_A(y) \cap X_k|/|\delta_A(y)|$;
4. $X_k = \arg\max\{\Pr(X_k|\delta_A(y)) : \forall X_k \in U/\mathrm{IND}_d\}$;
5. Find the corresponding decision label $\mathrm{Pre}_A(y)$ in terms of X_k;
6. **Return** $\mathrm{Pre}_A(y)$.

2.4 Attribute reduction

In general, the purpose of attribute reduction is to delete the redundant or irrelevant attributes, and then the rest can still meet the constraint. Based on such purpose, the attribute reduction with respect to the approximate quality can be defined as follows.

Definition 4 Given a decision system DS, $\forall A \subseteq AT$, A is referred to as a γ-reduct if and only if

1. $\gamma_A(d) \geq \gamma_{AT}(d)$;
2. $\forall B \subset A, \gamma_B(d) < \gamma_A(d)$.

Based on Definition 4, the approach to find reduct is not only worth to be addressed but also important. Up to now, the heuristic algorithm based on greedy strategy for finding reduct has been widely used because of its lower time consuming (Wang et al. 2016; Yang et al. 2019). In the framework of heuristic algorithm, the most significant attribute in each iteration is determined by a significance function (Yang and Yao 2018). For example, the significance function in terms of approximate quality is as follows.

Definition 5 Given a decision system DS, if $A \subset AT$, then $\forall a \in AT - A$, its significance with respect to approximate quality is :

$$\mathrm{Sig}_\gamma(a, A, d) = \gamma_{A\cup\{a\}}(d) - \gamma_A(d). \tag{8}$$

The detailed process of heuristic approach to compute reduct in terms of approximate quality is shown in Algorithm 2.

Algorithm 2. Compute γ-Reduct (CAQR)

Inputs: DS and radius σ;
Outputs: A γ-reduct A.
1. $A \leftarrow \emptyset$;
2. Compute $\gamma_{AT}(d)$;
3. **Do**
1) $\forall a_i \in AT - A$, compute $\mathrm{Sig}_\gamma(a_i, A, d)$;
`// Notes that` $\gamma_\emptyset(d) = 0$;
2) Select a_j such that $\mathrm{Sig}_\gamma(a_j, A, d) = \max\{\mathrm{Sig}_\gamma(a_i, A, d) : \forall a_i \in AT - A\}$;
3) $A \leftarrow A \cup \{a_j\}$;
4) Compute $\gamma_A(d)$;
Until $\gamma_A(d) \geq \gamma_{AT}(d)$;
4. **Return** A.

3 Attribute reduction for partially labeled data

Many significant functions such as the one used in Algorithm 2 are derived from labeled samples. It follows that such significant function can be employed to evaluate the significance of each attribute in terms of labeled samples. However, if unlabeled samples exist in data, then it is impossible for us to obtain the expected measurement. Moreover, if unlabeled samples are ignored, then the information given by these samples may be wasted. For such reason, attribute reduction we mentioned in the above section cannot be directly used for partially labeled data.

Hence, considering the importance of attribute based on both labeled and unlabeled samples, we will propose a combined importance. In such strategy, the neighborhood approximate quality is used to measure importance of attribute in terms of labeled samples while the neighborhood relation is employed to measure the importance of attribute in terms of unlabeled samples. The general process of constructing the new importance is shown in Fig. 1.

Through many experiments, it is trivial to observe that if the number of used attributes is increasing, then the value of approximate quality tends to be higher while the cardinality of neighborhood relation tends to be lower. Therefore, the higher the importance of an attribute, the greater the value of approximate quality and the finer the neighborhood relation. From this point of view, the new importance is defined in Definition 6.

Definition 6 Given a decision system DS, let $U = l \cup ul$, where l is the set of labeled samples and ul is the set of unlabeled samples, $\forall A \subseteq AT$, the importance of A is defined by:

$$\mathrm{IMP}(A) = \alpha \cdot \frac{\gamma_A^l(d)}{\gamma_{AT}^l(d)} + \beta \cdot \frac{|\delta_{AT}^{ul}|}{|\delta_A^{ul}|}, \tag{9}$$

where $\gamma_A^l(d)$, $\gamma_{AT}^l(d)$ are computed by labeled samples in l, $|\delta_{AT}^{ul}|, |\delta_A^{ul}|$ are obtained by unlabeled samples in ul. In addition, $\alpha + \beta = 1$ and their values are in [0, 1].

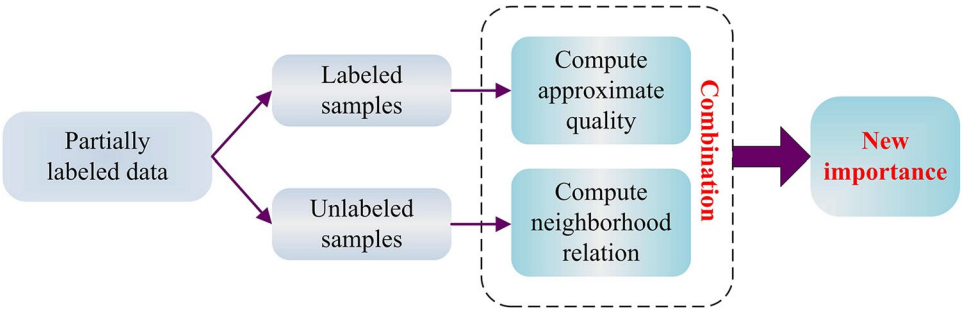

Fig. 1 Process of constructing new importance

Example 1 Let us consider the following example of partially labeled data with ten samples and four conditional attributes. Among these ten samples, seven samples are labeled and three samples are unlabeled.

Suppose that we want to compute the IMP($\{a_1\}$) in Table 1 by Definition 6 when the radius used in neighborhood is 0.1.

1. *For the first step of computation, we have*

$$\gamma^l_{AT}(d) = 0.2857 \quad and \quad |\delta^{ul}_{AT}| = 6.$$

2. *For the second step of computation, we have*

$$\gamma^l_{\{a_1\}}(d) = 0 \quad and \quad |\delta^{ul}_{\{a_1\}}| = 6.$$

Therefore, if α and β are set to 0.8 and 0.2, then $\mathrm{IMP}(\{a_1\}) = 0.8 \cdot (0/0.2857) + 0.2 \cdot (6/6) = 0.2$ is obtained by Eq. (9).

By Definition 6, the attribute reduction for partially labeled data with respect to IMP can be defined as follows.

Definition 7 Given a decision system DS, $\forall A \subseteq AT$, A is referred to as an IMP-reduct if and only if

1. $\mathrm{IMP}(A) \geq \mathrm{IMP}(AT)$;
2. $\forall B \subset A$, $\mathrm{IMP}(B) < \mathrm{IMP}(A)$.

Immediately, the significant function designed for IMP and the attribute reduction algorithm based on IMP can be constructed as follows. Note that to simplify our algorithm and reduce its time consumption, we focus on the suboptimal solution instead of the optimum solution (Yang et al. 2013, 2014). Therefore, the second condition in Definition 7 is not taken into account in this paper.

Definition 8 Given a decision system DS, if $A \subset AT$, then $\forall a \in AT - A$, its significance with respect to IMP is:

$$\mathrm{Sig}_{\mathrm{IMP}}(a, A) = \mathrm{IMP}(A \cup \{a\}) - \mathrm{IMP}(A). \tag{10}$$

Table 1 A small example of partially labeled data

Samples	a_1	a_2	a_3	a_4	d
1	0.8147	0.1576	0.6557	0.7061	1
2	0.9058	0.9706	0.0357	0.0318	1
3	0.1270	0.9572	0.8491	0.2769	2
4	0.9134	0.4854	0.9340	0.0462	3
5	0.6324	0.8003	0.6787	0.0971	2
6	0.0975	0.1419	0.7577	0.8235	3
7	0.2785	0.4218	0.7431	0.6948	1
8	0.5469	0.9157	0.3922	0.3171	?
9	0.9575	0.7922	0.6555	0.9502	?
10	0.9649	0.9595	0.1712	0.0344	?

Algorithm 3. Compute IMP-Reduct (CIMR)

Inputs: DS and radius σ;
Outputs: An IMP-reduct A.
1. $A \leftarrow \emptyset$;
2. Compute IMP(AT);
3. **Do**
  1) $\forall a_i \in AT - A$, compute $\mathrm{Sig}_{\mathrm{IMP}}(a_i, A)$;
    `// Notes that,` IMP($\emptyset$) = 0;
  2) Select a_j such that $\mathrm{Sig}_{\mathrm{IMP}}(a_j, A) = \max\{\mathrm{Sig}_{\mathrm{IMP}}(a_i, A) : \forall a_i \in AT - A\}$;
  3) $A \leftarrow A \cup \{a_j\}$;
  4) Compute IMP(A);
  Until IMP(A) $\geq$IMP(AT);
4. **Return** A.

4 Experiments

4.1 Datasets

To evaluate various performances of our CIMR, 12 real-world datasets from UCI machine learning repository have been employed in the following experiment. Table 2 summarizes some detailed statistics of these datasets used in our experiments.

4.2 Experimental setup

All the experiments have been carried out on a personal computer with Window7, Intel Core i5-3337U CPU (1.80 GHz) and 4.00 GB memory. The programming language is MATLAB R2014a.

In our experiments, tenfold cross-validation is employed for evaluating the effectiveness of different methods. And for each train set, we randomly divide it into two groups (groups of labeled and unlabeled samples) by ratios of 7:3, 5:5 and 3:7, i.e., three different ratios (30%, 50% and 70%) of missing labels. Moreover, each complete train set without missing labels (0% of missing labels) is retained for CAQR. And then we appoint ten different neighborhood radii such that $\sigma = 0.03, 0.06, \ldots, 0.3$.

To set α and β for a better performance of our approach, we conduct CIMR with different settings beforehand. Tenfold cross-validation is also employed, and then CIMR is performed over each train set. Through executing those derived reducts over each test set, the NEC based classification accuracies are compared. Note that for each setting of α and β, with such three ratios of missing labels, three reducts by CIMR can be derived from each train set. Correspondingly, three groups of classification accuracies can be obtained from each test set. And the averages of these

Table 2 Characteristics of the experimental datasets

ID	Datasets	Samples	Attributes	Decision classes
1	Breast Cancer Wisconsin (Diagnostic)	569	31	2
2	Breast Tissue	106	10	6
3	Climate Model Simulation Crashes	540	21	2
4	*Ecoli*	336	8	8
5	Hayes-Roth	132	5	3
6	Leaf	340	16	36
7	Seeds	210	8	3
8	Statlog (Image Segmentation)	2310	19	7
9	Steel Plates Faults	1941	34	2
10	Vertebral Column	310	7	2
11	Website Phishing	1353	10	2
12	Yeast	1484	9	10

classification accuracies are mainly compared as shown in Table 3. It is not difficult to observe that if α and β are set to be 0.8 and 0.2, then CIMR may be more effective. From this point of view, we set α and β to be 0.8 and 0.2 in our experiments.

4.3 Experimental results and analyses

In the following, with respect to three ratios (30%, 50% and 70%) of missing labels, the reducts derived by CAQR are denoted as 30%-AQR, 50%-AQR and 70%-AQR, respectively. Similar to such representation, the reducts obtained by CIMR are indicated as 30%-IMR, 50%-IMR and 70%-IMR, respectively. Note that CAQR is only executed over the labeled samples and the reduct derived by complete data with 0% of missing labels is denoted as 0%-AQR.

In our experiments, the lengths, values of approximate quality, values of conditional entropy and NEC based classification accuracies derived by reducts will be compared. The detailed experimental results are shown in the following.

With an investigation of Table 4, it is not difficult to observe: (1) with the same ratios of missing labels, IMRs are longer than AQRs; (2) if CAQR is executed over the complete data, the lengths of 0%-AQRs are still smaller. It is mainly because both the value of approximate quality derived by labeled samples and the neighborhood relation obtained by unlabeled samples should be considered synchronously, i.e., the constraint used in CIMR is stricter than that in CAQR.

By Fig. 2, it is not difficult to observe the following.

1. In general, the values of approximate quality derived by IMRs are higher than or equal to those by AQRs. As

Table 3 Classification accuracies among different settings of α and β (greater values are in bold)

ID	$\alpha = 0.1$ $\beta = 0.9$	$\alpha = 0.2$ $\beta = 0.8$	$\alpha = 0.3$ $\beta = 0.7$	$\alpha = 0.4$ $\beta = 0.6$	$\alpha = 0.5$ $\beta = 0.5$	$\alpha = 0.6$ $\beta = 0.4$	$\alpha = 0.7$ $\beta = 0.3$	$\alpha = 0.8$ $\beta = 0.2$	$\alpha = 0.9$ $\beta = 0.1$
1	0.9170	0.9162	0.9162	0.9117	0.9119	0.9160	0.9230	**0.9242**	0.9235
2	0.3873	0.4132	0.4511	0.4671	0.4795	0.4909	0.4855	**0.4911**	0.4875
3	0.9060	0.9058	0.9042	0.9028	0.9048	0.9030	0.9030	**0.9064**	0.9064
4	0.7921	0.7869	0.7869	0.7718	0.7782	0.7782	0.7754	**0.7921**	0.7611
5	0.5255	0.5489	0.5731	0.5868	0.5922	0.5992	0.6133	**0.6266**	0.6264
6	0.2082	0.2082	0.2082	0.2082	0.2082	0.2082	0.2082	**0.2082**	0.2082
7	0.8511	0.8438	0.8479	0.8457	0.8492	0.8486	0.8521	**0.8679**	0.8600
8	0.6908	0.7614	0.7946	0.7988	0.8014	0.8095	0.8096	**0.8244**	0.7897
9	0.8661	0.9108	0.9458	0.9748	0.9801	0.9909	**0.9942**	0.9925	0.9893
10	0.7234	0.7275	**0.7340**	0.7301	0.7262	0.7249	0.7219	0.7249	0.7249
11	0.7966	0.8006	0.8166	0.8406	0.8442	0.8450	0.8468	**0.8486**	0.8445
12	0.4215	0.4347	0.4311	0.4392	0.4523	0.4496	0.4554	**0.4574**	0.4554
Average	0.6738	0.6882	0.7008	0.7065	0.7107	0.7137	0.7157	**0.7203**	0.7148

Table 4 Comparisons among lengths of reducts (great values are in bold)

ID	30%-IMR	30%-AQR	50%-IMR	50%-AQR	70%-IMR	70%-AQR	0%-AQR
1	1.7	1.7	1.3	1.3	**1.6**	1.5	1.8
2	1.3	1.3	1.8	**2.0**	**1.9**	1.7	1.4
3	**3.9**	3.6	**4.8**	4.2	**4.1**	2.8	3.1
4	**3.1**	2.3	2.7	2.7	**2.8**	2.4	2.4
5	**2.9**	2.4	**2.4**	1.7	**3.1**	2.8	2.3
6	**4.9**	1.1	**1.9**	1.1	**12.7**	1.0	2.1
7	2.5	2.5	**2.3**	2.2	**1.3**	1.0	2.2
8	**4.7**	4.3	**4.4**	3.8	**4.7**	4.4	4.8
9	**8.6**	6.1	**9.6**	6.7	**9.3**	6.3	6.1
10	**2.1**	2.0	**1.6**	1.4	**1.3**	1.2	1.9
11	**5.5**	4.3	**5.3**	4.5	**5.1**	4.0	4.9
12	**3.9**	3.2	**2.3**	2.2	**4.4**	3.6	3.6
Average	**3.8**	2.9	**3.4**	2.8	**4.4**	2.7	3.1

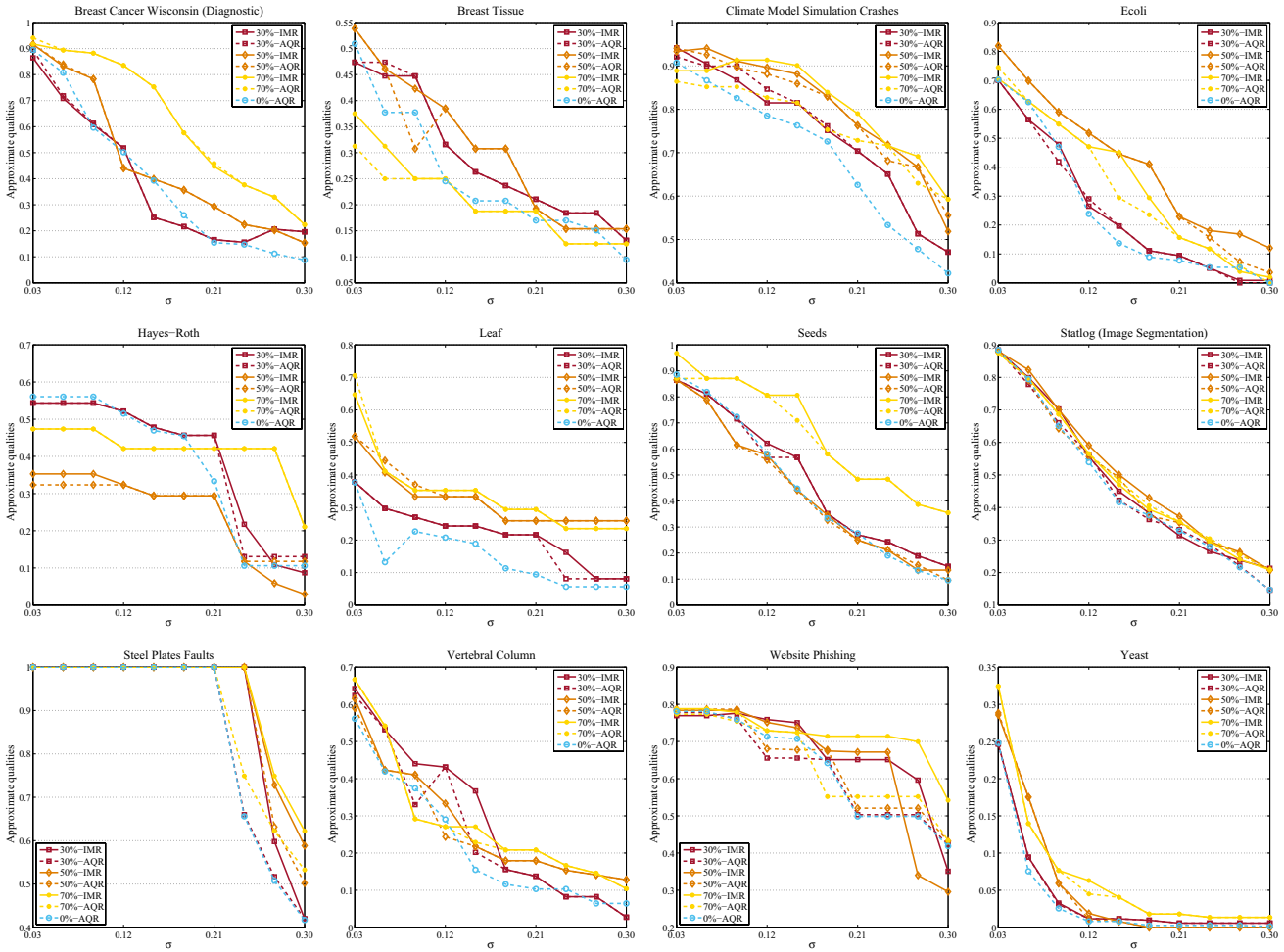

Fig. 2 Comparisons among values of approximate quality derived by reducts

shown in Table 4, IMRs are greater than AQRs. Therefore, we may conclude that a reduct with more attributes may bring us higher value of approximate quality.

2. If the radius is greater, then the derived value of approximate quality tends to be lower. It may because that with the greater radius, the value of approximate quality derived by raw data tends to be lower, which may lead to the looser constraints both in CAQR and CIMR. It follows that the corresponding reducts may offer us worse performances in improving the value of approximate quality.

By Fig. 3, the values of conditional entropy derived by IMRs are lower than or equal to those by AQRs. By Table 4, we may conclude that a reduct with more attributes may bring us lower value of conditional entropy. Moreover, similar to the results related to the values of approximate quality, it is easily to know that if the used radius is greater, then the value of conditional entropy derived by reduct may tend to be higher.

By Fig. 4, we can observe that with the same ratios of missing labels, the classification accuracies derived by IMRs are higher than or equal to those by AQRs. It may because in the process of computing reducts, CAQR ignores the unlabeled samples so that samples used in CIMR is more. Therefore, if the complete data are used, then the classification accuracies derived by 0%-AQRs may be higher.

With a thorough investigation of Figs. 2, 3 and 4, we can observe that from the viewpoint of used ratio, when the ratio of missing labels is 30%, the 30%-IMRs offer us the better performances. It may imply that although we have found an effective algorithm to deal with partially labeled data, if the ratio of missing labels is too high, then our algorithm will become much less effective.

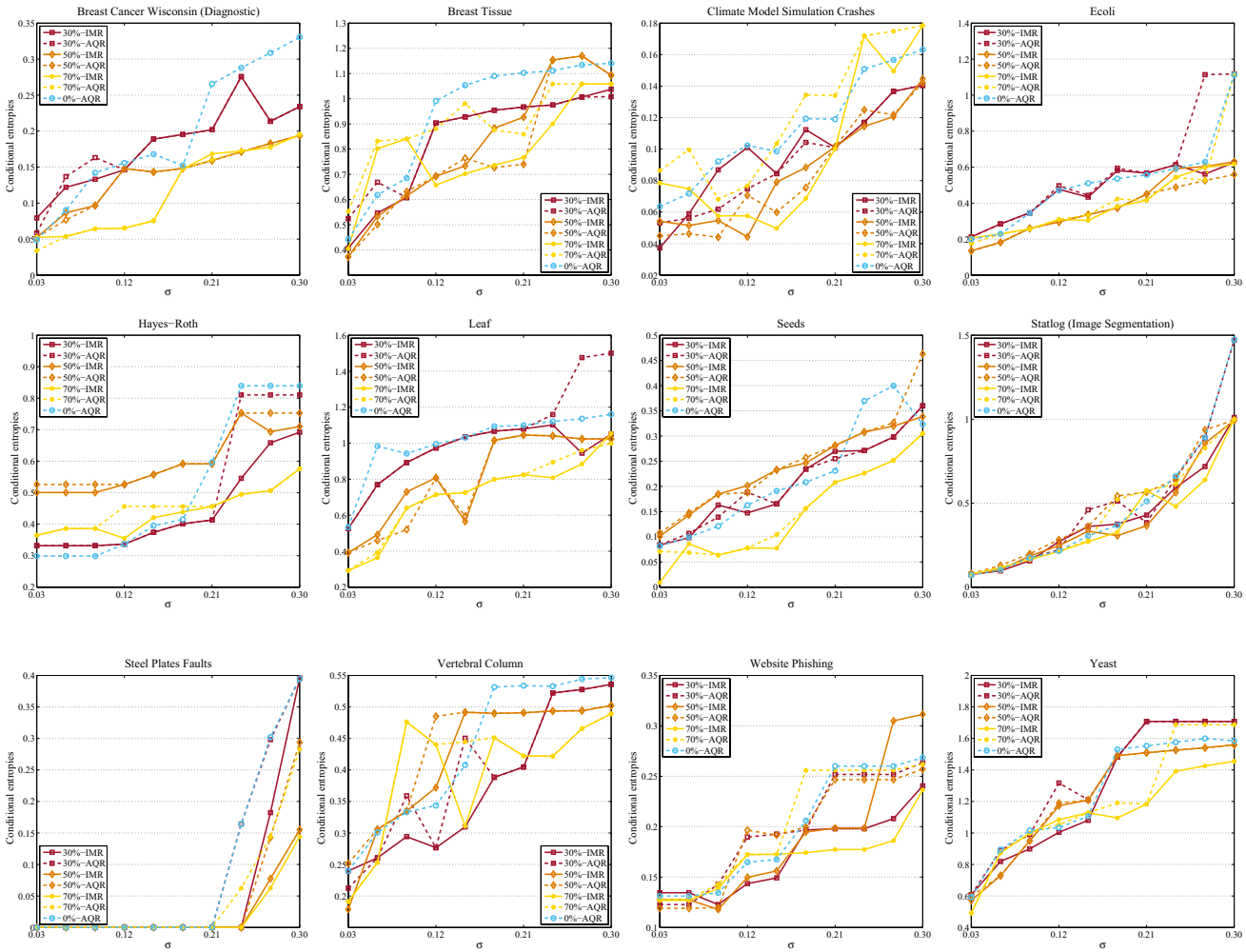

Fig. 3 Comparisons among the values of conditional entropy derived by reducts

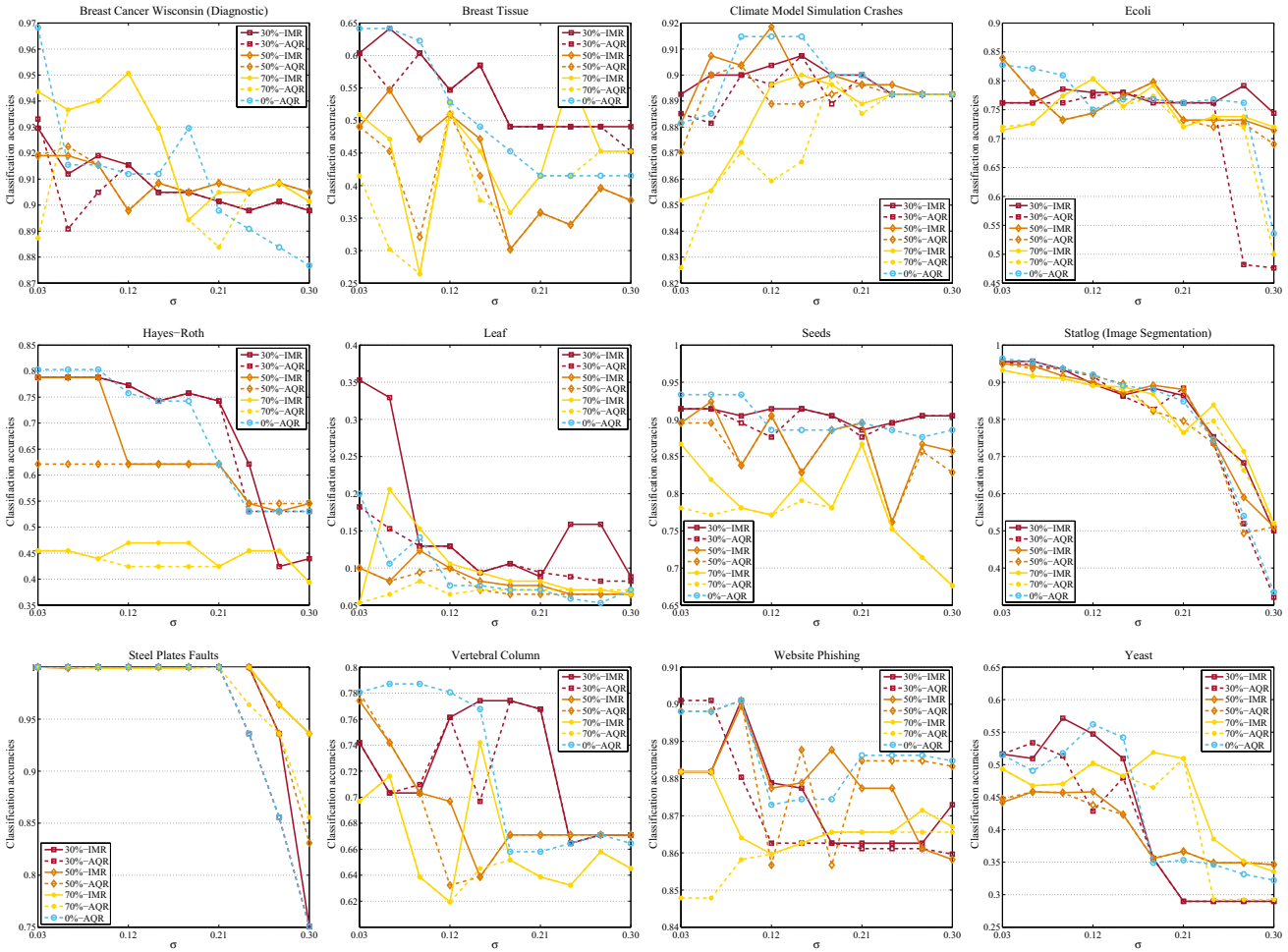

Fig. 4 Comparisons among classification accuracies derived by reducts

4.4 Statistics comparisons among the results

In the following, to further analyze the previous results from the viewpoint of statistics, the one-way analysis of variance (one-way ANOVA) (Fisher 1921) will be employed for comparing such results.

The one-way ANOVA can be used for comparing the averages of two or more groups of data. In the context of this paper, such method is used to rank the differences in performances of two different types of reducts. For example, for the ratios of missing labels 30%, with 10 different neighborhood radii, 10 different 30%-IMRs can be derived, then 10 values of approximate quality can be collected into a group. Similarly, 10 values of approximate quality in terms of 10 different 30%-AQRs can be collected into the other group. Immediately, the difference of the values of approximate quality in such two groups can be compared by the one-way ANOVA.

Note that the returned p value in the one-way ANOVA is under the null hypothesis that these groups are drawn from populations with the same average. If p value is higher than the default 5% significance level (0.05), it indicates that the averages of these groups are similar; instead, the averages of these groups are significantly different. The detailed results of p values are shown in Tables 5, 6 and 7.

Following the results shown in Tables 5, 6 and 7, it is not difficult to observe that in most cases, the returned p values are higher than 0.05. Therefore, we may conclude that the performances of IMRs derived by our approach cannot be worse than those of AQRs.

5 Conclusions and future perspectives

In this paper, we have designed a new algorithm of finding reduct for partially labeled data. The main contributions of this paper are: (1) a new importance for evaluating attribute is proposed, and such importance can be used for searching suitable attributes over partially labeled data; (2) through various comparative experiments, the final results indicate

Table 5 p values of one-way ANOVA for comparing the values of approximate quality derived by reducts

ID	30%-AQR& 30%-IMR	50%-AQR& 50%-IMR	70%-AQR& 70%-IMR	0%-AQR& 30%-IMR	0%-AQR& 50%-IMR	0%-AQR& 70%-IMR
1	0.9701	0.9956	0.9766	0.9590	0.6215	0.0853
2	0.9639	0.9516	0.7142	0.5070	0.4326	0.4411
3	0.9477	0.9220	0.2932	0.5046	0.1204	0.0776
4	0.9635	0.8594	0.8784	0.9777	0.1382	0.3989
5	0.9793	0.9078	1.0000	0.8327	0.0985	0.5750
6	0.4998	0.3951	0.7730	0.1345	0.1108	0.1104
7	0.9645	0.9620	0.8446	0.8143	0.9240	0.0841
8	0.8934	0.8414	0.9628	0.8777	0.6947	0.8142
9	0.6770	0.8069	0.5261	0.6676	0.4134	0.3702
10	0.7611	0.8722	0.9593	0.4731	0.4902	0.4417
11	0.3905	0.8852	0.0932	0.4859	0.7877	0.0849
12	0.9956	0.9905	0.9675	0.8809	0.6720	0.3941

Table 6 p values of one-way ANOVA for comparing the values of conditional entropy derived by reducts

ID	30%-AQR& 30%-IMR	50%-AQR& 50%-IMR	70%-AQR& 70%-IMR	0%-AQR& 30%-IMR	0%-AQR& 50%-IMR	0%-AQR& 70%-IMR
1	0.9280	0.9632	0.9222	0.6608	0.1100	0.0439
2	0.8208	0.7915	0.1882	0.3484	0.3252	0.1722
3	0.7511	0.8935	0.2550	0.2960	0.0825	0.4412
4	0.3330	0.7324	0.6372	0.6121	0.1946	0.1783
5	0.5234	0.6883	0.6155	0.4098	0.3597	0.3423
6	0.1044	0.6474	0.4808	0.0096	0.0596	0.0190
7	0.9780	0.7376	0.8663	0.8334	0.7008	0.1388
8	0.6058	0.6261	0.7023	0.6837	0.6660	0.5863
9	0.6609	0.5704	0.4098	0.6579	0.2244	0.2025
10	0.7378	0.6958	0.7716	0.3083	0.7599	0.4412
11	0.2428	0.8475	0.1278	0.2694	0.7431	0.1853
12	0.7498	0.9790	0.5874	0.8978	0.8899	0.3674

Table 7 p values of one-way ANOVA for comparing classification accuracies of reducts

ID	30%-AQR& 30%-IMR	50%-AQR& 50%-IMR	70%-AQR& 70%-IMR	0%-AQR& 30%-IMR	0%-AQR& 50%-IMR	0%-AQR& 70%-IMR
1	0.5290	0.9132	0.4463	0.8453	0.9030	0.2992
2	0.6131	0.5671	0.2090	0.2912	0.0721	0.1715
3	0.1476	0.1303	0.2883	0.8636	0.9419	0.0419
4	0.1369	0.8247	0.3878	0.6586	0.9838	0.7531
5	0.8575	0.1796	0.1857	1.0000	0.4420	0.0000
6	0.1465	0.4180	0.0662	0.0507	0.5800	0.7797
7	0.3077	0.7509	0.5249	0.4839	0.0544	0.0000
8	0.6341	0.7798	0.8080	0.7339	0.8249	0.7717
9	0.7005	0.4904	0.4071	0.7005	0.2189	0.2226
10	0.7222	0.7654	0.5499	0.9584	0.2022	0.0235
11	0.6442	0.3894	0.0243	0.0331	0.1194	0.0000
12	0.7277	0.9481	0.5023	0.7523	0.3726	0.6266

that our approach is effective in handling partially labeled data. The following topics deserve our further investigations.

1. We have only realized our algorithm with the concept of approximate quality in this paper, and some other measurements, such as conditional entropy and neighborhood decision error rate, will be further applied.
2. Attribute reduction can be considered as the previous step of data processing, and classification performances of different classifiers based on our reduct will be further explored.

Acknowledgements This work is supported by the National Natural Science Foundation of China (nos. 61572242, 61503160, 61502211), Macau Science and Technology Development Fund (no. 081/2015/A3).

References

Chen SM, Chang YC (2011) Weighted fuzzy rule interpolation based on GA-based weight-learning techniques. IEEE Trans Fuzzy Syst 19:729–744

Chen SM, Chen JH (2009) Fuzzy risk analysis based on ranking generalized fuzzy numbers with different heights and different spreads. Expert Syst Appl 36:6320–6334

Chen SM, Tanuwijaya K (2011) Fuzzy forecasting based on high-order fuzzy logical relationships and automatic clustering techniques. Expert Syst Appl 38:15425–15437

Chen SM, Lee SH, Lee CH (2001) A new method for generating fuzzy rules from numerical data for handling classification problems. Appl Artif Intell 15:645–664

Chen DG, Zhao SY, Zhang L, Yang YP, Zhang X (2012) Sample pair selection for attribute reduction with rough set. IEEE Trans Knowl Data Eng 24:2080–2093

Dai JH, Wang WT, Xu Q (2013) An uncertainty measure for incomplete decision tables and its applications. IEEE Trans Cybern 43:1277–1289

Dai JH, Hu QH, Zhang JH, Hu H, Zheng NG (2017) Attribute selection for partially labeled categorical data by rough set approach. IEEE Trans Cybern 47:2460–2471

Dou HL, Yang XB, Song XN, Yu HL, Wu WZ, Yang JY (2016) Decision-theoretic rough set: a multicost strategy. Knowl Based Syst 91:71–83

Dubois D, Prade H (1990) Rough fuzzy sets and fuzzy rough sets. Int J Gen Syst 17:191–209

Fisher RA (1921) On the "probable error" of a coefficient of correlation. Metron 1:3–32

Hu QH, Yu DR, Xie ZX, Liu JF (2006) Fuzzy probabilistic approximation spaces and their information measures. IEEE Trans Fuzzy Syst 14:191–201

Hu QH, Liu JF, Yu DR (2008a) Mixed feature selection based on granulation and approximation. Knowl Based Syst 21:294–304

Hu QH, Yu DR, Xie ZX (2008b) Neighborhood classifiers. Expert Syst Appl 34:866–876

Hu QH, Zhang L, Chen DG, Pedrycz W, Yu DR (2010) Gaussian kernel based fuzzy rough sets: model, uncertainty measures and applications. Int J Approx Reason 51:453–471

Hu J, Li TR, Wang HJ, Fujita H (2016) Hierarchical cluster ensemble model based on knowledge granulation. Knowl Based Syst 91:179–188

Huang B, Li HX (2018) Distance-based information granularity in neighborhood-based granular space. Granul Comput 3:93–110

Ju HR, Li HX, Yang XB, Zhou XZ, Huang B (2017) Cost-sensitive rough set: a multi-granulation approach. Knowl Based Syst 123:137–153

Liu KY, Yang XB, Yu HL, Mi JS, Wang PX, Chen XJ (2018) Rough set based semi-supervised feature selection via ensemble selector. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2018.11.034

Mi JS, Wu WZ, Zhang WX (2004) Approaches to knowledge reduction based on variable precision rough set model. Inf Sci 159:255–272

Min F, Xu J (2016) Semi-greedy heuristics for feature selection with test cost constraints. Granul Comput 1:199–211

Pawlak Z (1992) Rough sets: theoretical aspects of reasoning about data. Kluwer Academic Publishers, Dordrecht

Pawlak Z, Skowron A (2007) Rough sets: some extensions. Inf Sci 177:28–40

Pedrycz W, Chen SM (2011) Granular computing and intelligent systems: design with information granules of higher order and higher type. Springer, Heidelberg

Pedrycz W, Chen SM (2015) Granular computing and decision-making: interactive and iterative approaches. Springer, Heidelberg

Peter JF, Skowron A, Synak P, Ramanna S (2003) Rough sets and information granulation. In: Proceedings 10th international fuzzy systems association world congress, Istanbul, Turkey, pp 370–377

Polkowski L, Artiemjew P (2015) Granular computing in decision approximation: an application of rough mereology. Springer, Heidelberg

Skowron A, Stepaniuk J (1996) Tolerance approximation spaces. Fundamenta Informaticae 27:245–253

Swiniarski W, Skowron A (2003) Rough set methods in feature selection and recognition. Pattern Recognit Lett 24:83–849

Wang GY (2017) Dgcc: data-driven granular cognitive computing. Granul Comput 2:343–355

Wang HY, Chen SM (2008) Evaluating students' answerscripts using fuzzy numbers associated with degrees of confidence. IEEE Trans Fuzzy Syst 16:403–415

Wang CZ, Shao MW, He Q, Qian YH, Qi YL (2016) Feature subset selection based on fuzzy neighborhood rough sets. Knowl Based Syst 111:173–179

Wang GY, Yang J, Xu J (2017) Granular computing: from granularity optimization to multi-granularity joint problem solving. Granul Comput 2:105–120

Wang CZ, Hu QH, Wang XZ, Chen DG, Qian YH, Dong Z (2018) Feature selection based on neighborhood discrimination index. IEEE Trans Neural Netw Learn Syst 29:2986–2999

Wei W, Liang JY, Wang JH, Qian YH (2013) Decision-relative discernibility matrices in the sense of entropies. Int J Gen Syst 42:721–738

Wojna A (2005) Analogy-based reasoning in classifier construction. Trans Rough Sets IV 3700:277–374

Wu WZ, Qian YH, Li TJ, Gu SM (2016) On rule acquisition in incomplete multi-scale decision tables. Inf Sci 378:282–302

Xu SP, Yang XB, Yu HL, Yu DJ, Yang JY, Tsang ECC (2016) Multi-label learning with label-specific feature reduction. Knowl Based Syst 104:52–61

Xu WH, Li WT, Zhang XT (2017) Generalized multigranulation rough sets and optimal granularity selection. Granul Comput 2:271–288

Yang XB, Yao YY (2018) Ensemble selector for attribute reduction. Appl Soft Comput 70:1–11

Yang XB, Song XN, Dou HL, Yang JY (2011a) Multi-granulation rough set: from crisp to fuzzy case. Ann Fuzzy Math Inform 1:55–70

Yang XB, Zhang M, Dou HL, Yang JY (2011b) Neighborhood systems-based rough sets in incomplete information system. Knowl Based Syst 24:858–867

Yang XB, Qi YS, Song XN, Yang JY (2013) Test cost sensitive multigranulation rough set: model and minimal cost selection. Inf Sci 250:184–199

Yang XB, Qi Y, Yu HL, Song XN, Yang JY (2014) Updating multigranulation rough approximations with increasing of granular structures. Knowl Based Syst 64:59–69

Yang XB, Liang SC, Yu HL, Gao S, Qian YH (2019) Pseudo-label neighborhood rough set: measures and attribute reductions. Int J Approx Reason 105:112–129

Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353

Zhang X, Mei CL, Chen DG, Li JH (2016) Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy. Pattern Recognit 56:1–15

Zhi HL, Li JH (2018) Granule description based on positive and negative attributes. Granul Comput 3:1–14

Zhou ZH, Li M (2010) Semi-supervised learning by disagreement. Knowl Inf Syst 24:415–439

Zhu P, Wen QY (2012) Information-theoretic measures associated with rough set approximations. Inf Sci 212:33–43