International Journal of Approximate Reasoning 105 (2019) 112–129

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Pseudo-label neighborhood rough set: Measures and attribute reductions [☆]

Xibei Yang [a,b,*], Shaochen Liang [a], Hualong Yu [a], Shang Gao [a], Yuhua Qian [c,d]

[a] School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212003, PR China
[b] School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, PR China
[c] School of Computer and Information Technology, Shanxi University, Taiyuan 030006, PR China
[d] Intelligent Information Processing Key Laboratory of Shanxi Province, Shanxi University, Taiyuan 030006, PR China

ARTICLE INFO

Article history:
Received 13 March 2018
Received in revised form 8 November 2018
Accepted 13 November 2018
Available online 19 November 2018

Keywords:
Attribute reduction
Conditional discrimination index
Conditional entropy
Neighborhood decision error rate
Neighborhood rough set
Pseudo-label

ABSTRACT

The scale of the radius for constructing neighborhood relation has a great effect on the results of neighborhood rough sets and corresponding measures. A very small radius frequently brings us nothing because any two different samples are separated from each other, though these two samples have the same label. If the radius is growing, then there is a serious risk that samples with different labels may fall into the same neighborhood. Obviously, the radius based neighborhood relation does not take the labels of samples into account, which will lead to unsatisfactory discrimination. To fill such gap, a pseudo-label strategy is systematically studied in rough set theory. Firstly, a pseudo-label neighborhood relation is proposed. Such relation can differentiate samples by not only the distance but also the pseudo labels of samples. Therefore, both the neighborhood rough set and some corresponding measures can be re-defined. Secondly, attribute reductions are explored based on the re-defined measures. The heuristic algorithm is also designed to compute reducts. Finally, the experimental results over UCI data sets tell us that our pseudo-label strategy is superior to the traditional neighborhood approach. This is mainly because the former can significantly reduce the uncertainties and improve the classification accuracies. The Wilcoxon signed rank test results also show that neighborhood approach and pseudo-label neighborhood approach are so different from the viewpoints of the measures and attribute reductions in rough set theory.

1. Introduction

Up to now, neighborhood rough set has witnessed a great success in the development of rough set theory due to the following reasons. Firstly, the neighborhood relation derived from the distance function provides us a valuable framework for analyzing data with continuous or even mixed values [10,16]. Secondly, samples can be separated from each other by adopting different radii, it follows that different granularities for discriminating samples can be obtained, i.e., the structure of multi-granularity [3,4,20,21,24,27,59,61,62] is naturally obtained. Finally, neighborhood rough set is an effective tool for incremental learning tasks [19,28,33,39,46,63], this is mainly because the neighborhood can be updated by using both the

[☆] This paper is part of the Virtual special issue on Uncertainty in Granular Computing, Edited by Duoqian Miao and Yiyu Yao.

[*] Corresponding author at: School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212003, PR China.

E-mail addresses: jsjxy_yxb@just.edu.cn (X. Yang), lshc940302@163.com (S. Liang), yuhualong@just.edu.cn (H. Yu), gao_shang@just.edu.cn (S. Gao), jinchengqyh@126.com (Y. Qian).

https://doi.org/10.1016/j.ijar.2018.11.010
0888-613X/© 2018 Elsevier Inc. All rights reserved.

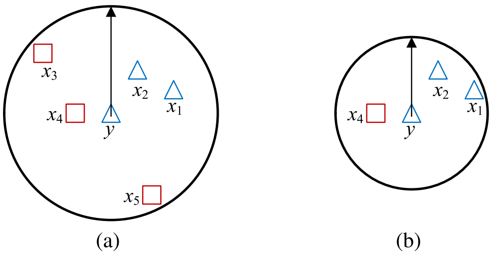

Fig. 1. Two neighborhoods with different radii. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

raw neighborhood and the updated part of the data without re-scanning the whole data [9,35,50,65]. In the light of these merits, neighborhood rough set has been successfully applied to different fields [2,22,23,26,30,55,67].

Generally speaking, the size of the neighborhood is a key factor in neighborhood rough set. The commonly used radii and distances have immediate effects on such size. For example, given a radius, if it is less than the distance between two samples, then these two samples are considered as distinguishable; if such radius is greater than or equal to the distance between two samples, then these two samples will fall into the same neighborhood, they are regarded as undistinguishable. From this point of view, the radius and distance will determine the number of samples in neighborhood and then the size of neighborhood.

Nevertheless, the above approach may have some inherent limitations. On the one hand, though smaller radius may result in a higher discrimination, it is very possible that two samples with the same label have been separated from each other. This is mainly because two samples with same label do not always have higher similarity in some practical applications. On the other hand, if the radius becomes greater, then intuitively, there is a serious risk that samples with different labels fall into the same neighborhood. The reason is that the lower quality of the conditional attributes (features) may result in a higher similarity between two samples with different labels. For such reasons, we can observe that the traditional way to construct neighborhood does not take the label information of samples into consideration.

Take the following Fig. 1 as an example. Obviously, the radius in the sub-figure (b) is smaller than that in sub-figure (a).

- In sub-figure (b), by the definition of rough set, sample y does not belong to the lower approximation of the class of triangle. This is mainly because: 1) x_4 is very close to y; 2) x_4 and y have different labels. It is a typical inconsistent case. Moreover, if the value of radius is increased as sub-figure (a) shows, then more samples whose labels are different from the label of y will fall into the neighborhood of y, e.g., x_3 and x_5. The inconsistent case also holds. This example tells us that if the labels of samples are not taken into consideration, then it is difficult to derive better approximation.
- Moreover, in sub-figure (b), y can be correctly classified by the neighborhood classifier [13,56] if the majority rule is employed. Furthermore, if the radius value keeps reducing, then it is possible that only x_4 is in the neighborhood of y if the reflexivity is ignored, in such case, y will be misclassified. This is mainly because x_4 is the nearest neighbor of y. Such observation indicates that the computation of distance between y and x_4 does not pay much attention to the labels of these two samples and then the misclassification will happen.

From discussions above, we can see that the labels of samples may affect the immediate results of neighborhoods. Therefore, a mechanism which contains the information provided by the labels is desired. Though the decision attribute used in rough set theory offers us the labels of samples, these labels cannot be directly used. This is mainly because the labels in decision attribute are used to derive decision classes for approximation. It is unreasonable to approximate the decision classes generated by decision attribute by using the label information provided by the decision attribute itself. Fortunately, motivated by the research results of pseudo-label strategy in unsupervised and semi-supervised learning tasks [5,31,41–43,49,64,70], we know that the pseudo labels of samples provide us with another information of labels. From this point of view, it is a useful attempt to introduce the pseudo-label strategy into neighborhood rough set theory, and this is what will be mainly addressed in this paper.

From the viewpoint of Granular Computing, the neighborhood relation offers us a mechanism of information granulation. It must be noticed that if the pseudo labels of samples are considered, then another type of information granulation can also be executed. Consequently, more than one result of information granulation have been obtained. As it is pointed out in References [32,38], two techniques can be employed for rough set data analysis if more than one type of information granulation is considered: 1) fuse the results of these information granulations and then construct rough set; 2) construct different rough sets based on different results of information granulations and then obtain multigranulation rough sets [34]. In this paper, we will adopt the former technique. The reason can be attributed to two aspects: firstly, the motivation of our pseudo-label strategy is to improve the discrimination of neighborhood relation rather than multigranulation rough set; secondly, since the derived pseudo labels of samples are discrete instead of continuous, pseudo-label based information granulation will generate a partition in which the number of equivalence classes is equivalent to the number of real labels, it follows that the lower approximations based on such partition may be very small, such result is meaningless for constructing multigranulation rough sets.

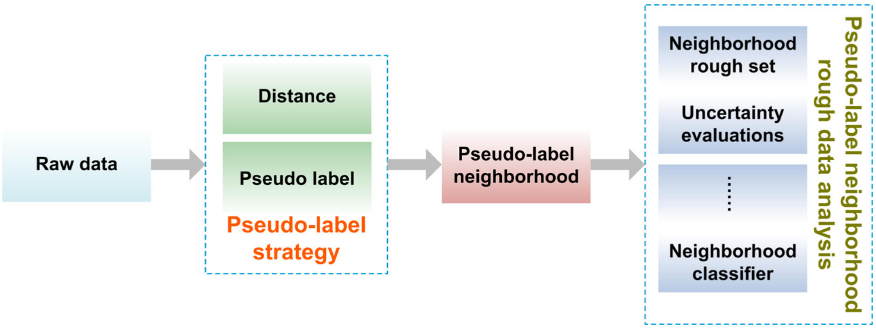

Fig. 2. The framework of this research.

The following Fig. 2 reports the framework of our research. Obviously, the main contribution of our approach is the pseudo-label strategy. In other words, not only the distance, but also the pseudo labels are used for constructing the neighborhoods of samples. Immediately, all concepts and approaches related to neighborhood rough set can be transformed into the pseudo-label neighborhood based case.

The rest of the paper are organized as follows. Basic notions related to neighborhood rough set are introduced in Section 2. Pseudo-label neighborhood rough set and the corresponding attribute reductions are studied in Section 3. The experimental results and comparisons are discussed in Section 4. Section 5 is the conclusion of the paper.

2. Preliminary knowledge

2.1. Neighborhood relation

Without loss of generality, a neighborhood decision system can be represented as $\mathrm{NDS} =< U, A, d >$ in which U is the set of samples, A is the set of condition attributes and d is a decision attribute. $\forall x \in U$, $d(x)$ indicates the label of sample x, and $a(x)$ denotes its value over condition attribute $a \in A$.

Given a neighborhood decision system, 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 $U/\mathrm{IND}_d = \{X_1, X_2, \ldots, X_q\}$ is induced. In rough set theory, $X_k \in U/\mathrm{IND}_d$ is called the k-th decision class. Especially, $\forall x \in U$, the decision class which contains sample x is denoted by $[x]_d$.

Moreover, a relation can also be defined in terms of condition attributes. Since the values of most of the condition attributes are continuous in real applications, then $\forall B \subseteq A$, Hu et al. [13] have defined a neighborhood relation such that $\mathrm{N}_B = \{(x, y) \in U \times U : \Delta_B(x, y) \le \sigma\}$. In N_B, σ is the radius such that $\sigma \ge 0$, and $\Delta_B(\cdot, \cdot)$ is the distance function [48] with respect to B, which satisfies the following properties.

1) Non-negativity: $\forall x, y \in U$, $\Delta_B(x, y) \ge 0$, and $\Delta_B(x, y) = 0$ if and only if $x = y$.
2) Symmetry: $\forall x, y \in U$, $\Delta_B(x, y) = \Delta_B(y, x)$.
3) Triangle inequality: $\forall x, y, z \in U$, $\Delta_B(x, z) \le \Delta_B(x, y) + \Delta_B(y, z)$.

In the context of this paper, Euclidean distance is employed, i.e., $\Delta_B(x, y) = \sqrt{\sum_{a \in B} \left(a(x) - a(y)\right)^2}$. By N_B, the neighborhood of sample x is formed such that $\mathrm{N}_B(x) = \{y \in U : (x, y) \in \mathrm{N}_B\}$. To avoid sample x from being the only sample in the neighborhood of x, which may bring us difficulty for constructing neighborhood classifier, Hu et al. [13] have modified the radius of σ for each $x \in U$ such that

$$\delta_x = \min_{y \in U - \{x\}} \left\{\Delta_B(x, y)\right\} + \sigma \cdot \left(\max_{y \in U - \{x\}} \left\{\Delta_B(x, y)\right\} - \min_{y \in U - \{x\}} \left\{\Delta_B(x, y)\right\} \right). \tag{1}$$

Following Equation (1), the updated neighborhood relation is denoted by δ_B such that $\delta_B = \{(x, y) \in U \times U : \Delta_B(x, y) \le \delta_x\}$, then the neighborhood of x is $\delta_B(x) = \{y \in U : \Delta_B(x, y) \le \delta_x\}$.

2.2. Neighborhood rough set

Following the neighborhood we mentioned above, the lower and upper approximations can be defined as follows.

Definition 1. [13] Given a neighborhood decision system NDS, $U/\mathrm{IND}_d = \{X_1, X_2, \ldots, X_q\}$, $\forall B \subseteq A$, the neighborhood lower and upper approximations of d with respect to B are

Table 1
An example of neighborhood decision system.

	a_1	a_2	a_3	a_4	a_5	d
x_1	0.2518	0.9827	0.9063	0.0225	0.4229	1
x_2	0.2904	0.7302	0.8797	0.4253	0.0942	1
x_3	0.6171	0.3439	0.8178	0.3127	0.5985	2
x_4	0.2653	0.5841	0.2607	0.1615	0.4709	2
x_5	0.8244	0.1078	0.5944	0.1788	0.6959	2

$$\underline{\delta_B}(d) = \bigcup_{k=1}^{q} \underline{\delta_B}(X_k), \tag{2}$$

$$\overline{\delta_B}(d) = \bigcup_{k=1}^{q} \overline{\delta_B}(X_k), \tag{3}$$

where $\forall X_k \in U/\mathrm{IND}_d$,

$$\underline{\delta_B}(X_k) = \{x \in U : \delta_B(x) \subseteq X_k\}, \tag{4}$$

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

The pair $\left[\underline{\delta_B}(X_k), \overline{\delta_B}(X_k)\right]$ is called a neighborhood rough set of X_k.

2.3. Some measures in neighborhood decision system

Approximation quality is a frequently used measure for describing the degree of certain belongingness in rough set theory. Following Definition 1, the corresponding approximation quality is defined as follows.

Definition 2. [13] Given a neighborhood decision system NDS, $\forall B \subseteq A$, the approximation quality of d with respect to B is defined as

$$\gamma_B(d) = \frac{|\underline{\delta_B}(d)|}{|U|}, \tag{6}$$

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

Approximation quality reflects the percentage of the samples which belong to one of the decision classes by the explanation of neighborhood lower approximation. Therefore, the higher the value of the approximation quality, the higher the degree of certain belongingness. Obviously, $0 \leq \gamma_B(d) \leq 1$ holds.

Furthermore, through a large number of experiments, it should be noticed that $\forall B \subset A$, $\gamma_B(d) \leq \gamma_A(d)$ does not always hold. The main reason is that with the variations of the used attributes, the modified radius δ obtained by Equation (1) will also change. An example is presented as follows.

Example 1. Let $\mathrm{NDS} =< U, A, d >$ be a neighborhood decision system shown in Table 1, where $U = \{x_1, x_2, x_3, x_4, x_5\}$, $A = \{a_1, a_2, a_3, a_4, a_5\}$ and $d(x) \in \{1, 2\}$. Let radius σ be 0.3.

By Equation (1), if A is used, then the modified radii δ for the five samples are 0.7466, 0.7313, 0.5382, 0.7614 and 0.6344, respectively. Consequently, it is obtained that $\gamma_A(d) = 0.8$.

Suppose that the condition attribute a_1 is deleted from the neighborhood decision system and then let $B = \{a_2, a_3, a_4, a_5\}$, following Equation (1), the modified radii δ for the five samples are 0.7001, 0.6892, 0.5123, 0.7455 and 0.5879, respectively. Therefore, $\gamma_B(d) = 1$. Obviously, $\gamma_B(d) > \gamma_A(d)$.

Conditional entropy is another measure which characterizes the discriminating ability of $B \subseteq A$ relative to d. Generally speaking, the higher the discrimination of $B \subseteq A$ relative to d, the lower the uncertainty in the neighborhood decision system. Presently, with respect to different requirements, many different definitions of conditional entropies have been proposed [6,7,14,15,25,68,69]. A widely used form of conditional entropy is shown in the following definition.

Definition 3. [66] Given a neighborhood decision system NDS, $\forall B \subseteq A$, the conditional entropy of d with respect to B is defined as

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

Algorithm 1 Neighborhood Classifier (NEC).

```
Inputs: NDS $=< U, A, d >$, $B \subseteq A$, test sample $y \notin U$ and radius $\sigma$;
Outputs: Predicted label $\mathrm{Pre}_B(y)$.
1. $\forall x \in U$, compute $\Delta_B(y, x)$;
2. Compute modified radius $\delta_y$, and obtain $\delta_B(y)$;
3. Compute $U/\mathrm{IND}_d$;
4. $\forall X_k \in U/\mathrm{IND}_d$, compute the probability $\Pr(X_k|\delta_B(y)) = \frac{|\delta_B(y) \cap X_k|}{|\delta_B(y)|}$;
5. $X_i = \arg\max\{\Pr(X_k|\delta_B(y)) : \forall X_k \in U/\mathrm{IND}_d\}$;
6. Find the corresponding label $\mathrm{Pre}_B(y)$ in terms of decision class $X_i$;
7. Return $\mathrm{Pre}_B(y)$.
```

It can be proved that $0 \leq \mathrm{ENT}_B(d) \leq |U|/e$ holds [66]. Similar to what has been addressed for approximation quality, $\forall B \subset A, \mathrm{ENT}_B(d) \geq \mathrm{ENT}_A(d)$ does not always hold.

By Equation (7), we can see that the value of conditional entropy can be obtained if and only if the neighborhoods of all samples have been obtained. The process of deriving neighborhoods by neighborhood relation is very time-consuming because the time complexity is $O(|U|^2)$. Therefore, Wang et al. [44] have proposed the concept of conditional discrimination index which can be directly obtained by the neighborhood relation instead of neighborhoods.

Definition 4. [44] Given a neighborhood decision system NDS, $\forall B \subseteq A$, the conditional discrimination index of d with respect to B is defined as

$$\mathrm{H}_B(d) = \log \frac{|\delta_B|}{|\delta_B \cap \mathrm{IND}_d|}. \tag{8}$$

The lower the value of the conditional discrimination index, the higher the consistent degree of the neighborhood decision system. Obviously, $0 \leq \mathrm{H}_B(d) \leq \log |U|$ holds. For example, if δ_B is an identity relation, then $\mathrm{H}_B(d)$ achieves the minimal value 0. The identity relation indicates that any two samples in U can be separated from each other by B, it follows that $\delta_B \subseteq \mathrm{IND}_d$ and then NDS is completely consistent. Another extreme example, if $\delta_B = \{(x, y) : \forall x, y \in U\}$ and IND_d is the identity relation, then $\mathrm{H}_B(d)$ achieves the maximal value $\log |U|$. In such case, $\delta_B \nsubseteq \mathrm{IND}_d$ and then NDS is completely inconsistent. It should be noticed that in Reference [44], Wang et al. have pointed out that conditional discrimination index is not monotonic, i.e., $\forall B \subset A, \mathrm{H}_B(d) \geq \mathrm{H}_A(d)$ does not always hold.

The above measures are all defined from the viewpoints of the certainty or uncertainty in neighborhood decision system. Nevertheless, the classification performance of neighborhood approach is also worthy to be investigated. Therefore, Hu et al. [11] have proposed a measure called neighborhood decision error rate in neighborhood decision system. Such measure can be obtained by the neighborhood classifier shown in Algorithm 1.

Based on the neighborhood classifier, neighborhood decision error rate is defined as follows.

Definition 5. [11] Given a neighborhood decision system NDS, $\forall B \subseteq A$, the neighborhood decision error rate of d with respect to B is defined as

$$\mathrm{NDER}_B(d) = \frac{|\{x \in U : \mathrm{Pre}_B(x) \neq d(x)\}|}{|U|}. \tag{9}$$

For each computation of $\mathrm{Pre}_B(x)$ in Equation (9), x is considered as a test sample and used as an input in Algorithm 1. If the predicted label of x is obtained, then it can be compared with the true label of x. This process implies that the neighborhood decision error rate is actually generated by a leave-one-out validation strategy. Therefore, $\mathrm{NDER}_B(d)$ reflects the percentage of the misclassified samples. Obviously, $0 \leq \mathrm{NDER}_B(d) \leq 1$ holds.

2.4. Attribute reduction in neighborhood decision system

Since the above four measures are not always monotonic with the variations of used condition attributes, then by these measures, we can re-define the corresponding criteria for attribute reductions. Different from traditional attribute reductions [11,17,52,53] in rough set theory, these criteria aim to decrease the uncertainties or improve the classification performance instead of preserving them.

Definition 6. Given a neighborhood decision system NDS, $\forall B \subseteq A$,

1. B is referred to as an Approximation Quality-reduct(γ-reduct) if and only if $\gamma_B(d) \geq \gamma_A(d)$ and $\forall C \subset B$, $\gamma_C(d) < \gamma_A(d)$;
2. B is referred to as a Conditional Entropy-reduct(CE-reduct) if and only if $\mathrm{ENT}_B(d) \leq \mathrm{ENT}_A(d)$ and $\forall C \subset B$, $\mathrm{ENT}_C(d) > \mathrm{ENT}_A(d)$;

3. B is referred to as a Conditional Discrimination Index-reduct(CDI-reduct) if and only if $\mathrm{H}_B(d) \le \mathrm{H}_A(d)$ and $\forall C \subset B$, $\mathrm{H}_C(d) > \mathrm{H}_A(d)$;
4. B is referred to as a Neighborhood Decision Error Rate-reduct(NDER-reduct) if and only if $\mathrm{NDER}_B(d) \le \mathrm{NDER}_A(d)$ and $\forall C \subset B$, $\mathrm{NDER}_C(d) > \mathrm{NDER}_A(d)$.

Following Definition 6, we can see that a γ-reduct is actually a minimal subset of A, which will not contribute to a lesser value of the approximation quality. Similarly, the other three reducts are also minimal subsets of A, which will not increase the values of the corresponding measures, respectively. It should be emphasized that though the explanations of these reducts are slightly different, these reducts can be derived by a similar heuristic algorithm, see References [1,12,18,36,37,40,45,47,51,54,57,58] for more details about heuristic algorithm used in rough set.

3. Pseudo-label neighborhood decision system

By the neighborhood relation shown in the above section, we can clearly observe that different radii σ will result in different levels of discriminations. The smaller the value of σ is, it is possible that more samples can be separated from each other. However, according to what has been discussed in Section 1, the information provided by labels of samples has been ignored in deriving neighborhood relation. Therefore, it is possible that two samples with different labels may fall into the same neighborhood, which will make it difficult in generating satisfactory approximations. For such reason, to improve the discrimination ability of neighborhood relation, the labels of samples should be taken into account. Furthermore, it should be noticed that the real labels of samples cannot be used directly, it is mainly because we cannot approximate the decision classes generated by decision attribute by using the label information provided by the decision attribute itself. Therefore, in Section 3.1, a pseudo-label neighborhood strategy will be proposed. Such pseudo-label strategy is realized by using a pseudo-label attribute which is different from the decision attribute in the traditional decision system.

3.1. Pseudo-label neighborhood rough set

Formally, a pseudo-label neighborhood decision system can be represented as a generalization of the neighborhood decision system such that $\mathrm{NDS}^{\mathrm{PL}} =< U, A, d, d_A^{\mathrm{PL}} >$, in which d_A^{PL} is referred to as the pseudo-label attribute. $\forall x \in U$, $d_A^{\mathrm{PL}}(x)$ expresses the pseudo label of x, which can be derived from a learning approach based on A, and the learning approaches may be clustering analysis, classification analysis or the label propagation [29], etc.

Since both condition attributes and pseudo labels of samples exist in the pseudo-label neighborhood decision system, we can define the following pseudo-label neighborhood relation for replacing the traditional neighborhood relation shown in Section 2.1.

Definition 7. Given a pseudo-label neighborhood decision system $\mathrm{NDS}^{\mathrm{PL}}$, $\forall B \subseteq A$, the pseudo-label neighborhood relation is defined such that

$$\delta_B^{\mathrm{PL}} = \{(x, y) \in U \times U : \Delta_B(x, y) \le \delta_x \wedge d_B^{\mathrm{PL}}(x) = d_B^{\mathrm{PL}}(y)\}, \tag{10}$$

in which δ_x is the modified radius shown in Equation (1).

The pseudo-label neighborhood relation shown in the above definition tells us that two samples are regarded as indistinguishable if and only if: 1) their distance are less than or equal to the radius; 2) such two samples should have the same pseudo label.

Following Definition 7, the pseudo-label neighborhood of x is then defined as

$$\delta_B^{\mathrm{PL}}(x) = \{y \in U : \Delta_B(x, y) \le \delta_x \wedge d_B^{\mathrm{PL}}(x) = d_B^{\mathrm{PL}}(y)\}. \tag{11}$$

Proposition 1. *Given a pseudo-label neighborhood decision system* $\mathrm{NDS}^{\mathrm{PL}}$, $\forall B \subseteq A$, *we have* $\delta_B^{\mathrm{PL}} \subseteq \delta_B$ *and* $\delta_B^{\mathrm{PL}} = \delta_B \cap \mathrm{IND}_{d_B^{\mathrm{PL}}}$ *where* $\mathrm{IND}_{d_B^{\mathrm{PL}}} = \{(x, y) \in U \times U : d_B^{\mathrm{PL}}(x) = d_B^{\mathrm{PL}}(y)\}$*.*

Proof. It can be derived directly by the forms of δ_B^{PL} and δ_B. □

The above proposition tells us that by the pseudo-label strategy, we can obtain a finer neighborhood relation which provides higher performance of discrimination.

Proposition 2. *Given a pseudo-label neighborhood decision system* $\mathrm{NDS}^{\mathrm{PL}}$, $\forall B \subseteq A$ *and* $\forall x \in U$, *we have*

$$\delta_B^{\mathrm{PL}}(x) \subseteq \delta_B(x). \tag{12}$$

Proof. It can be derived directly by the definitions of $\delta_B^{\text{PL}}(x)$ and $\delta_B(x)$. □

Immediately, a pseudo-label neighborhood rough set can be defined as follows.

Definition 8. Given a pseudo-label neighborhood decision system NDS^{PL}, $U/\text{IND}_d = \{X_1, X_2, \ldots, X_q\}$, $\forall B \subseteq A$, the pseudo-label neighborhood lower and upper approximations of d with respect to B are

$$\underline{\delta_B^{\text{PL}}}(d) = \bigcup_{k=1}^{q} \underline{\delta_B^{\text{PL}}}(X_k), \tag{13}$$

$$\overline{\delta_B^{\text{PL}}}(d) = \bigcup_{k=1}^{q} \overline{\delta_B^{\text{PL}}}(X_k), \tag{14}$$

where $\forall X_k \in U/\text{IND}_d$,

$$\underline{\delta_B^{\text{PL}}}(X_k) = \{x \in U : \delta_B^{\text{PL}}(x) \subseteq X_k\}, \tag{15}$$

$$\overline{\delta_B^{\text{PL}}}(X_k) = \{x \in U : \delta_B^{\text{PL}}(x) \cap X_k \neq \emptyset\}. \tag{16}$$

The pair $[\underline{\delta_B^{\text{PL}}}(X_k), \overline{\delta_B^{\text{PL}}}(X_k)]$ is referred to as a pseudo-label neighborhood rough set of X_k.

Proposition 3. *Given a pseudo-label neighborhood decision system* NDS^{PL}, $\forall X_k \in U/\text{IND}_d$, *we have* $\underline{\delta_B}(X_k) \subseteq \underline{\delta_B^{\text{PL}}}(X_k)$ *where* $\delta > 0$ *is the given radius.*

Proof. $\forall x \in \underline{\delta_B}(X_k)$, by Equation (4), we have $\delta_B(x) \subseteq X_k$. Moreover, by the result of Proposition 2, we know that $\delta_B^{\text{PL}}(x) \subseteq \delta_B(x)$, it follows that $\delta_B^{\text{PL}}(x) \subseteq X_k$ and then $x \in \underline{\delta_B^{\text{PL}}}(X_k)$ holds, i.e., $\underline{\delta_B}(X_k) \subseteq \underline{\delta_B^{\text{PL}}}(X_k)$. □

3.2. Some measures in pseudo-label neighborhood decision system

Similar to what we have discussed in neighborhood decision system, the four measures can also be defined in pseudo-label neighborhood decision system. The details will be addressed as follows.

Definition 9. Given a pseudo-label neighborhood decision system NDS^{PL}, $\forall B \subseteq A$, the pseudo-label approximation quality of d with respect to B is defined as

$$\gamma_B^{\text{PL}}(d) = \frac{|\underline{\delta_B^{\text{PL}}}(d)|}{|U|}. \tag{17}$$

The higher the value of the pseudo-label approximation quality, the higher the degree of certain belongingness in pseudo-label neighborhood decision system. Obviously, $0 \leq \gamma_B^{\text{PL}}(d) \leq 1$ holds. If $\underline{\delta_B^{\text{PL}}}(d) = \emptyset$, then pseudo-label approximation quality will achieve the minimal value 0; if $\underline{\delta_B^{\text{PL}}}(d) = U$, then pseudo-label approximation quality will achieve the maximal value 1.

Moreover, $\forall B \subset A$, $\gamma_B^{\text{PL}}(d) \leq \gamma_A^{\text{PL}}(d)$ does not always hold. There are two reasons. On the one hand, with the variations of the used attributes, the modified radius δ obtained by Equation (1) will change. On the other hand, different pseudo-labels of samples will be derived from a given learning approach when different attributes are used. An example is presented as follows.

Example 2. Let $\text{NDS} =< U, A, d >$ be a pseudo-label neighborhood decision system shown in Table 2, where $U = \{x_1, x_2, x_3, x_4, x_5\}$, $A = \{a_1, a_2, a_3, a_4, a_5\}$, $d(x) \in \{1, 2\}$, d_A^{PL} is the pseudo-label decision attribute such that $d_A^{\text{PL}} \in \{\star, *\}$. Let radius σ be 0.6.

By Equation (1), if A is used, then the modified radii δ for the five samples are 0.8653, 0.9351, 0.7665, 0.7751 and 0.8255, respectively. Therefore, it is obtained that $\gamma_A^{\text{PL}}(d) = 0.2$.

Assuming that the condition attribute a_1 is deleted from the neighborhood decision system and then $B = \{a_2, a_3, a_4, a_5\}$. It should be emphasized that the pseudo labels of the five samples should be re-derived based on B. Let's say the obtained pseudo labels are $\star$, $*$, $*$, $*$ and $*$, respectively. Moreover, by Equation (1), the modified radii δ for the five samples are 0.8194, 0.8892, 0.6652, 0.7258 and 0.7742, respectively. Consequently, $\gamma_B^{\text{PL}}(d) = 0.6$. Obviously, $\gamma_B^{\text{PL}}(d) > \gamma_A^{\text{PL}}(d)$ holds.

Proposition 4. *Given a pseudo-label neighborhood decision system* NDS^{PL}, $\forall B \subseteq A$, *we have* $\gamma_B^{\text{PL}}(d) \geq \gamma_B(d)$.

Table 2
An example of pseudo-label neighborhood decision system.

	a_1	a_2	a_3	a_4	a_5	d	d_A^{PL}
x_1	0.6401	0.4230	0.5055	0.0323	0.2277	1	$\star$
x_2	0.2360	0.1091	0.2719	0.7036	0.9019	1	$\star$
x_3	0.7737	0.0333	0.6002	0.2800	0.3518	2	$\star$
x_4	0.2458	0.2427	0.7494	0.3817	0.5664	2	$\ast$
x_5	0.6148	0.4514	0.1644	0.6846	0.1889	2	$\star$

Proof. By Proposition 3, we have $\underline{\delta_B}(X_k) \subseteq \underline{\delta_B^{\mathrm{PL}}}(X_k)$. Therefore, by Equations (2) and (13), we obtain $\underline{\delta_B^{\mathrm{PL}}}(d) \supseteq \underline{\delta_B}(d)$. Following Definitions 2 and 9, it is obvious that $\gamma_B^{\mathrm{PL}}(d) \geq \gamma_B(d)$ holds. □

The above proposition implies that the pseudo-label strategy will bring us higher approximation quality when compared with the traditional neighborhood rough set approach.

Definition 10. Given a pseudo-label neighborhood decision system $\mathrm{NDS}^{\mathrm{PL}}$, $\forall B \subseteq A$, the pseudo-label conditional entropy of d with respect to B is defined as

$$\mathrm{ENT}_B^{\mathrm{PL}}(d) = -\frac{1}{|U|}\sum_{x\in U}\left|\delta_B^{\mathrm{PL}}(x)\cap[x]_d\right|\log\frac{|\delta_B^{\mathrm{PL}}(x)\cap[x]_d|}{|\delta_B^{\mathrm{PL}}(x)|}. \tag{18}$$

The lower the value of the pseudo-label conditional entropy, the higher the certainty degree of the pseudo-label neighborhood decision system.

Proposition 5. *Given a pseudo-label neighborhood decision system* $\mathrm{NDS}^{\mathrm{PL}}$*,* $\forall B \subseteq A$*, we have* $\mathrm{ENT}_B^{\mathrm{PL}}(d) \leq \mathrm{ENT}_B(d)$*.*

Proof. Note that $|\delta_B(x)| = |\delta_B(x)\cap[x]_d| + |\delta_B(x)\cap(U-[x]_d)|$ and $|\delta_B^{\mathrm{PL}}(x)| = |\delta_B^{\mathrm{PL}}(x)\cap[x]_d| + |\delta_B^{\mathrm{PL}}(x)\cap(U-[x]_d)|$. Let $|\delta_B(x)\cap[x]_d| = m_1^x$, $|\delta_B(x)\cap(U-[x]_d)| = m_2^x$, $|\delta_B^{\mathrm{PL}}(x)\cap[x]_d| = n_1^x$ and $|\delta_B^{\mathrm{PL}}(x)\cap(U-[x]_d)| = n_2^x$. Therefore,

$$\mathrm{ENT}_B(d) = -\frac{1}{|U|}\sum_{x\in U} m_1^x \log\frac{m_1^x}{m_1^x+m_2^x};$$

$$\mathrm{ENT}_B^{\mathrm{PL}}(d) = -\frac{1}{|U|}\sum_{x\in U} n_1^x \log\frac{n_1^x}{n_1^x+n_2^x}.$$

Let $f(s,t) = -s\log\frac{s}{s+t}$ where $s>0$, $t\geq 0$ and $r = \frac{s}{s+t}$, it follows that $\frac{\partial f}{\partial s} = -\frac{t}{s+t} - \log\frac{s}{s+t} = -1 + r - \log r$ and $\frac{\partial f}{\partial t} = \frac{s}{s+t} > 0$. Since $(-1+r-\log r)' = 1-\frac{1}{r}$ and $r\in(0,1]$, then $\frac{\partial f}{\partial s} > 0$ also holds. Therefore, the function $f(\cdot,\cdot)$ is increasing with respect to the first and second arguments, respectively.

By Proposition 2, we have $\delta_B^{\mathrm{PL}}(x) \subseteq \delta_B(x)$ for each $x\in U$, it follows that $n_1^x \leq m_1^x$ and $n_2^x \leq m_2^x$, which yields $f(n_1^x, n_2^x) \leq f(m_1^x, m_2^x)$. Therefore, $\sum_{x\in U} f(n_1^x,n_2^x) \leq \sum_{x\in U} f(m_1^x,m_2^x)$, i.e., $\mathrm{ENT}_B^{\mathrm{PL}}(d) \leq \mathrm{ENT}_B(d)$ holds. □

The above proposition implies that by the pseudo-label based neighborhood, lower conditional entropy will be achieved when compared with the traditional neighborhood approach.

Proposition 6. *Given a pseudo-label neighborhood decision system* $\mathrm{NDS}^{\mathrm{PL}}$*,* $\forall B \subseteq A$*, we have* $0 \leq \mathrm{ENT}_B^{\mathrm{PL}}(d) \leq |U|/e$*.*

Proof. Firstly, let us prove that if $\forall x\in U$, $\delta_B^{\mathrm{PL}}(x) = U$ and $[x]_d = |U|/e$, then $\mathrm{ENT}_B^{\mathrm{PL}}(d)$ achieves the maximal value.

Suppose that there is a $x\in U$ and $\delta_B^{\mathrm{PL}}(x) \neq U$, the obtained pseudo-label conditional entropy is greater than the above maximal value, we then have $-\frac{1}{|U|}\sum_{x\in U}|\delta_B^{\mathrm{PL}}(x)\cap[x]_d|\log\frac{|\delta_B^{\mathrm{PL}}(x)\cap[x]_d|}{|\delta_B^{\mathrm{PL}}(x)|} > -\frac{1}{|U|}\sum_{x\in U}|U\cap[x]_d|\log\frac{|U\cap[x]_d|}{|U|} = -\frac{1}{|U|}\sum_{x\in U}|[x]_d|\log\frac{|[x]_d|}{|U|}$.

Similar to what have been addressed in Proposition 5, since the function $f(s,t) = -s\log\frac{s}{s+t}(s>0, t\geq 0)$ is increasing with respect to the first and second arguments, respectively, then we know that $-\frac{1}{|U|}\sum_{x\in U}|\delta_B^{\mathrm{PL}}(x)\cap[x]_d|\log\frac{|\delta_B^{\mathrm{PL}}(x)\cap[x]_d|}{|\delta_B^{\mathrm{PL}}(x)|} < -\frac{1}{|U|}\sum_{x\in U}|[x]_d|\log\frac{|[x]_d|}{|U|}$ because by the assumption, $\exists x\in U$ such that $\delta_B^{\mathrm{PL}}(x)\cap[x]_d \subset [x]_d$ and $\delta_B^{\mathrm{PL}}(x)\subset U$. This result contradicts the assumption.

Following the above proof, $\forall x \in U$, if $\delta_B^{\mathrm{PL}}(x) = U$, then $\mathrm{ENT}_B^{\mathrm{PL}}(d) = -\frac{1}{|U|}\sum_{x\in U}|[x]_d|\log\frac{|[x]_d|}{|U|}$. Let $f(s) = s\log\frac{s}{|U|}$, then $f'(s) = 1 + \log\frac{s}{|U|}$ and $f''(s) = \frac{1}{s} > 0$. If $s_0 = |U|/e$, then $f'(s_0) = 0$ implies $f(s_0) = \min f(s) = -|U|/e$. Therefore, if $[x]_d = |U|/e$ for each $x \in U$, then the obtained maximal value of $\mathrm{ENT}_B^{\mathrm{PL}}(d)$ is $|U|/e$.

By Proposition 5, since the function $f(s,t) = -s\log\dfrac{s}{s+t}(s > 0, t \geq 0)$ is increasing with respect to the first and second arguments, respectively, then $f(s,t)$ achieves the minimal value if both s and t are minimal values. Since pseudo-label neighborhood relation is at least reflexive, then the minimal value of $|\delta_B^{\mathrm{PL}}(x) \cap [x]_d|$ is 1 when $\delta_B^{\mathrm{PL}}(x) = \{x\}$, correspondingly, $|\delta_B^{\mathrm{PL}}(x) \cap (U - [x]_d)|$ achieves the minimal value 0. In other words, if $\delta_B^{\mathrm{PL}}(x) = \{x\}$ for each $x \in U$, then the minimal value of $\mathrm{ENT}_B^{\mathrm{PL}}(d)$ is 0. □

Definition 11. Given a pseudo-label neighborhood decision system $\mathrm{NDS}^{\mathrm{PL}}$, $\forall B \subseteq A$, the pseudo-label conditional discrimination index of d with respect to B is defined as

$$\mathrm{H}_B^{\mathrm{PL}}(d) = \log\frac{|\delta_B^{\mathrm{PL}}|}{|\delta_B^{\mathrm{PL}} \cap \mathrm{IND}_d|}. \tag{19}$$

The lower the value of the pseudo-label conditional discrimination index, the higher the degree of certainty in neighborhood decision system.

Proposition 7. *Given a pseudo-label neighborhood decision system* $\mathrm{NDS}^{\mathrm{PL}}$, $\forall B \subseteq A$, *we have* $\mathrm{H}_B^{\mathrm{PL}}(d) \leq \mathrm{H}_B(d)$.

Proof. By Equation (8), Equation (19) and Proposition 1, we have

$$\begin{aligned}\mathrm{H}_B^{\mathrm{PL}}(d) - \mathrm{H}_B(d) &= \log\frac{|\delta_B^{\mathrm{PL}}|}{|\delta_B^{\mathrm{PL}} \cap \mathrm{IND}_d|} - \log\frac{|\delta_B|}{|\delta_B \cap \mathrm{IND}_d|}\\ &= \log\frac{|\delta_B \cap \mathrm{IND}_{d^{\mathrm{PL}}}|}{|\delta_B \cap \mathrm{IND}_{d^{\mathrm{PL}}} \cap \mathrm{IND}_d|} - \log\frac{|\delta_B|}{|\delta_B \cap \mathrm{IND}_d|}\\ &= \log\frac{|\delta_B \cap \mathrm{IND}_{d^{\mathrm{PL}}}|}{|\delta_B \cap \mathrm{IND}_{d^{\mathrm{PL}}} \cap \mathrm{IND}_d|} \cdot \frac{|\delta_B \cap \mathrm{IND}_d|}{|\delta_B|}.\end{aligned}$$

Therefore, to prove $\mathrm{H}_B^{\mathrm{PL}}(d) \leq \mathrm{H}_B(d)$, it should be proved that $\log\dfrac{|\delta_B \cap \mathrm{IND}_{d^{\mathrm{PL}}}|}{|\delta_B \cap \mathrm{IND}_{d^{\mathrm{PL}}} \cap \mathrm{IND}_d|} \cdot \dfrac{|\delta_B \cap \mathrm{IND}_d|}{|\delta_B|} \leq 0$, i.e., $\dfrac{|\delta_B \cap \mathrm{IND}_{d^{\mathrm{PL}}}| \cdot |\delta_B \cap \mathrm{IND}_d|}{|\delta_B \cap \mathrm{IND}_{d^{\mathrm{PL}}} \cap \mathrm{IND}_d| \cdot |\delta_B|} \leq 1$. Moreover, δ_B, $\mathrm{IND}_{d^{\mathrm{PL}}}$ and IND_d considered in this paper are at least reflexive, it follows that $\dfrac{|\delta_B \cap \mathrm{IND}_{d^{\mathrm{PL}}}| \cdot |\delta_B \cap \mathrm{IND}_d|}{|\delta_B \cap \mathrm{IND}_{d^{\mathrm{PL}}} \cap \mathrm{IND}_d| \cdot |\delta_B|} \leq \dfrac{|\delta_B \cap \mathrm{IND}_{d^{\mathrm{PL}}}| \cdot |\delta_B \cap \mathrm{IND}_d|}{|\delta_B \cap \mathrm{IND}_{d^{\mathrm{PL}}} \cap \mathrm{IND}_d| \cdot |U|}$. Here, $|\delta_B| = |U|$ indicates that δ_B is an identity relation and then $\dfrac{|\delta_B \cap \mathrm{IND}_{d^{\mathrm{PL}}}| \cdot |\delta_B \cap \mathrm{IND}_d|}{|\delta_B \cap \mathrm{IND}_{d^{\mathrm{PL}}} \cap \mathrm{IND}_d| \cdot |\delta_B|} \leq \dfrac{|U| \cdot |U|}{|U| \cdot |U|} = 1$.

To sum up, the proposition is proved. □

The above proposition tells us that by comparing with the traditional neighborhood approach, the pseudo-label based neighborhood will generate the lower conditional discrimination index.

Proposition 8. *Given a pseudo-label neighborhood decision system* $\mathrm{NDS}^{\mathrm{PL}}$, $\forall B \subseteq A$, *we have* $0 \leq \mathrm{H}_B^{\mathrm{PL}}(d) \leq \log|U|$.

Proof. Obviously, $\delta_B^{\mathrm{PL}} \cap \mathrm{IND}_d \subseteq \delta_B^{\mathrm{PL}}$ holds and then we have $\dfrac{|\delta_B^{\mathrm{PL}}|}{|\delta_B^{\mathrm{PL}} \cap \mathrm{IND}_d|} \geq 1$. Immediately, the minimal value of $\log\dfrac{|\delta_B^{\mathrm{PL}}|}{|\delta_B^{\mathrm{PL}} \cap \mathrm{IND}_d|}$ is 0 when $\delta_B^{\mathrm{PL}} \subseteq \mathrm{IND}_d$.

Moreover, if $\log\dfrac{|\delta_B^{\mathrm{PL}}|}{|\delta_B^{\mathrm{PL}} \cap \mathrm{IND}_d|}$ wants to achieve the maximal value, then $\dfrac{|\delta_B^{\mathrm{PL}}|}{|\delta_B^{\mathrm{PL}} \cap \mathrm{IND}_d|}$ should be the maximal one, i.e., $|\delta_B^{\mathrm{PL}}|$ should achieve the maximal value and $|\delta_B^{\mathrm{PL}} \cap \mathrm{IND}_d|$ should achieve the minimal value. Obviously, $|\delta_B^{\mathrm{PL}}|$ will be the maximal one if and only if $\delta_B^{\mathrm{PL}} = U \times U$. Therefore, when $\delta_B^{\mathrm{PL}} = U \times U$ holds, the minimal value of $|\delta_B^{\mathrm{PL}} \cap \mathrm{IND}_d|$ will be achieved if and only if $\mathrm{IND}_d = \{(x, y) \in U \times U : x = y\}$. Consequently, the maximal value of $\mathrm{H}_B^{\mathrm{PL}}(d)$ is $\log|U|$. □

In Section 2.3, it has been pointed out that the concept of the neighborhood decision error rate can be used to characterize the classification performance of neighborhood classifier. Similarly, the pseudo-label neighborhood decision error

Algorithm 2 Pseudo-label Neighborhood Classifier (PNEC).

```
Inputs: $\mathrm{NDS}^{\mathrm{PL}} =< U, A, d, d^{\mathrm{PL}} >$, $B \subseteq A$, test sample $y \notin U$ and radius $\sigma$;
Outputs: Predicted label $\mathrm{Pre}_B^{\mathrm{PL}}(y)$.
$\forall x \in U$, compute $\Delta_B(y, x)$;
Compute modified radius $\delta$ and obtain pseudo-label neighborhood $\delta_B^{\mathrm{PL}}(y)$ by Equation (11);
$\forall X_k \in U/\mathrm{IND}_d$, compute the probability $\Pr(X_k|\delta_B^{\mathrm{PL}}(y)) = \frac{|\delta_B^{\mathrm{PL}}(y) \cap X_k|}{|\delta_B^{\mathrm{PL}}(y)|}$;
$X_i = \arg\max\{\Pr(X_k|\delta_B^{\mathrm{PL}}(y)) : \forall X_k \in U/\mathrm{IND}_d\}$;
Find the corresponding $\mathrm{Pre}_B^{\mathrm{PL}}(y)$ in terms of $X_i$;
Return $\mathrm{Pre}_B^{\mathrm{PL}}(y)$.
```

rate can also be defined. To achieve that, the neighborhood classifier should be redesigned. Different from the traditional neighborhood classifier, pseudo-label neighborhood classifier uses the pseudo-label neighborhood instead of the traditional neighborhood. Nevertheless, the majority principle is still used in pseudo-label neighborhood classifier. The detailed process of pseudo-label neighborhood classifier is shown in the following Algorithm 2.

Based on Algorithm 2, the classification performance of pseudo-label neighborhood classifier can be evaluated as follows.

Definition 12. Given a pseudo-label neighborhood decision system $\mathrm{NDS}^{\mathrm{PL}}$, $\forall B \subseteq A$, the pseudo-label neighborhood decision error rate of d with respect to B is defined as

$$\mathrm{NDER}_B^{\mathrm{PL}}(d) = \frac{|\{x \in U : \mathrm{Pre}_B^{\mathrm{PL}}(x) \neq d(x)\}|}{|U|}. \tag{20}$$

For each computation of $\mathrm{Pre}_B^{\mathrm{PL}}(x)$ in Equation (20), x is considered as a test sample and used as an input of Algorithm 2. Then the predicted label of x is obtained, and it can be compared with the true label of x. Obviously, $0 \leq \mathrm{NDER}_B^{\mathrm{PL}}(d) \leq 1$ also holds. However, different from the previous three measures, $\mathrm{NDER}_B^{\mathrm{PL}}(d) \leq \mathrm{NDER}_B(d)$ does not always hold.

3.3. Attribute reduction in pseudo-label neighborhood decision system

By the above measures defined in the pseudo-label neighborhood decision system, attribute reductions can be redefined as follows.

Definition 13. Given a pseudo-label neighborhood decision system $\mathrm{NDS}^{\mathrm{PL}}$, $\forall B \subseteq A$,

1. B is a Pseudo-label Approximation Quality-reduct(PL-γ-reduct) if and only if $\gamma_B^{\mathrm{PL}}(d) \geq \gamma_A^{\mathrm{PL}}(d)$ and $\forall C \subset B$, $\gamma_C^{\mathrm{PL}}(d) < \gamma_A^{\mathrm{PL}}(d)$;
2. B is a Pseudo-label Conditional Entropy-reduct(PLCE-reduct) if and only if $\mathrm{ENT}_B^{\mathrm{PL}}(d) \leq \mathrm{ENT}_A^{\mathrm{PL}}(d)$ and $\forall C \subset B$, $\mathrm{ENT}_C^{\mathrm{PL}}(d) > \mathrm{ENT}_A^{\mathrm{PL}}(d)$;
3. B is a Pseudo-label Conditional Discrimination Index-reduct(PLCDI-reduct) if and only if $\mathrm{H}_B^{\mathrm{PL}}(d) \leq \mathrm{H}_A^{\mathrm{PL}}(d)$ and $\forall C \subset B$, $\mathrm{H}_C^{\mathrm{PL}}(d) > \mathrm{H}_A^{\mathrm{PL}}(d)$;
4. B is a Pseudo-label Neighborhood Decision Error Rate-reduct(PLNDER-reduct) if and only if $\mathrm{NDER}_B^{\mathrm{PL}}(d) \leq \mathrm{NDER}_A^{\mathrm{PL}}(d)$ and $\forall C \subset B$, $\mathrm{NDER}_C^{\mathrm{PL}}(d) > \mathrm{NDER}_A^{\mathrm{PL}}(d)$.

To compute the above four pseudo-label reducts by heuristic algorithm, the significance functions are separately designed as follows.

Definition 14. Given a pseudo-label neighborhood decision system $\mathrm{NDS}^{\mathrm{PL}}$, if $B \subset A$, then $\forall a \in A - B$, its significances with respect to four different pseudo-label measures are:

$$\mathrm{Sig}_{\gamma}^{\mathrm{PL}}(a, B, d) = \gamma_{B \cup \{a\}}^{\mathrm{PL}}(d) - \gamma_B^{\mathrm{PL}}(d); \tag{21}$$

$$\mathrm{Sig}_{\mathrm{ENT}}^{\mathrm{PL}}(a, B, d) = \mathrm{ENT}_B^{\mathrm{PL}}(d) - \mathrm{ENT}_{B \cup \{a\}}^{\mathrm{PL}}(d); \tag{22}$$

$$\mathrm{Sig}_{\mathrm{H}}^{\mathrm{PL}}(a, B, d) = \mathrm{H}_B^{\mathrm{PL}}(d) - \mathrm{H}_{B \cup \{a\}}^{\mathrm{PL}}(d); \tag{23}$$

$$\mathrm{Sig}_{\mathrm{NDER}}^{\mathrm{PL}}(a, B, d) = \mathrm{NDER}_B^{\mathrm{PL}}(d) - \mathrm{NDER}_{B \cup \{a\}}^{\mathrm{PL}}(d). \tag{24}$$

In pseudo-label neighborhood decision system, all the significance functions presented above satisfy that the higher the value, the more significant the condition attribute a. For example, if $\mathrm{Sig}_{\mathrm{NDER}}^{\mathrm{PL}}(a_1, B, d) > \mathrm{Sig}_{\mathrm{NDER}}^{\mathrm{PL}}(a_2, B, d)$ where $a_1, a_2 \in A - B$, then we have $\mathrm{NDER}_{B \cup \{a_1\}}^{\mathrm{PL}}(d) < \mathrm{NDER}_{B \cup \{a_2\}}^{\mathrm{PL}}(d)$, such result indicates that compared with a_2, if a_1 is added into B, then the derived pseudo-label neighborhood decision error rate will be much lower.

Algorithm 3 Heuristic Algorithm for Computing PL-γ-reduct.

```
Inputs: $NDS^{PL} =< U, A, d, d^{PL} >$, and radius $\sigma$;
Outputs: A PL-$\gamma$-reduct $B$.
1.  $B \leftarrow \emptyset$;
2.  Compute modified radius $\delta$ and obtain pseudo-label approximation quality $\gamma_A^{PL}(d)$ by Equation (11);
3.  Do
        1) $\forall a \in A - B$, derive pseudo-label attribute $d_{B\cup\{a\}}^{PL}$;
        2) Compute $Sig_\gamma^{PL}(a, B, d)$; // $\gamma_\emptyset^{PL}(d) = 0$;
        3) Select $b$ such that $Sig_\gamma^{PL}(b, B, d) = \max\{Sig_\gamma^{PL}(a, B, d) : \forall a \in A - B\}$;
        4) $B \leftarrow B \cup \{b\}$;
        5) Compute $\gamma_B^{PL}(d)$
    Until $\gamma_B^{PL}(d) \geq \gamma_A^{PL}(d)$;
4.  Return $B$.
```

Table 3
Data sets description.

ID	Data set	Number of samples	Number of attributes	Number of decision classes
1	Breast Tissue	106	9	6
2	Cardiotocography	2126	21	10
3	Dermatology	366	34	6
4	Ecoli	336	7	8
5	Forest Type Mapping	523	27	4
6	Glass Identification	214	9	6
7	Libras Movement	360	90	15
8	Parkinsons	195	23	7
9	Statlog (Image Segmentation)	2310	18	7
10	Statlog (Vehicle Silhouettes)	846	18	4
11	Wine	178	13	3
12	Yeast	1484	8	10

Take the PL-γ-reduct as an example, the following Algorithm 3 will find such reduct by the significance function shown in Equation (21).

In Algorithm 3, note that the pseudo label of each sample should be re-derived for each iteration. Similarly, it is not difficult to revise Algorithm 3 for generating the other three reducts by using Equations (22), (23) and (24), respectively.

To facilitate the discussion of the time complexity of Algorithm 3, assuming that the pseudo labels of samples are derived from k-means clustering. Firstly, $Sig_\gamma^{PL}(a, B, d)$ is computed at most $(1 + |A|)|A|/2$ times in Algorithm 3. Secondly, the producing of pseudo labels requires extra time. Assume that the number of clusters is K and the iteration times of k-means is T, then the time complexity of producing pseudo labels is $O(KT|U||A|^3)$. Such time complexity is based on the two facts: 1) the time complexity of k-means is $O(KT|U||A|)$; 2) pseudo labels are produced $(1 + |A|)|A|/2$ times. Finally, the time complexity of Algorithm 3 is $O(|U|^2|A|^3 + KT|U||A|^3)$.

4. Experiments

4.1. Data sets

To verify the effectiveness of our pseudo-label strategy, 12 UCI data sets have been selected to conduct the experiments. Their details are displayed in Table 3. Note that the "Number of attributes" column refers to the number of condition attributes. For all experiments in this section, 10 different σ have been selected, they are $\sigma = 0.1, 0.2, \cdots, 1.0$.

4.2. Experimental results and experimental analyses

In this section, two groups of comparative experiments have been designed. Both of them are conducted on a personal computer with Intel i7-6700HP CPU (2.60 GHz) and 8 GB memory. In the experiments, k-means clustering is employed to produce pseudo labels, and the value of k is same as the number of decision classes in data.

4.2.1. Comparisons of the performances with respect to some measures

In this experiment, we will compare the performances of the neighborhood decision system and the pseudo-label neighborhood decision system with respect to the four measures considered in this paper. Figs. 3 and 4 report the detailed results of comparisons.

In Fig. 3, "γ" expresses the approximation quality, while the pseudo-label approximation quality is denoted by "PL-γ"; "CE" is represented as the conditional entropy, and "PLCE" represents the pseudo-label conditional entropy.

With a careful investigation of Fig. 3, it is not difficult to observe the following.

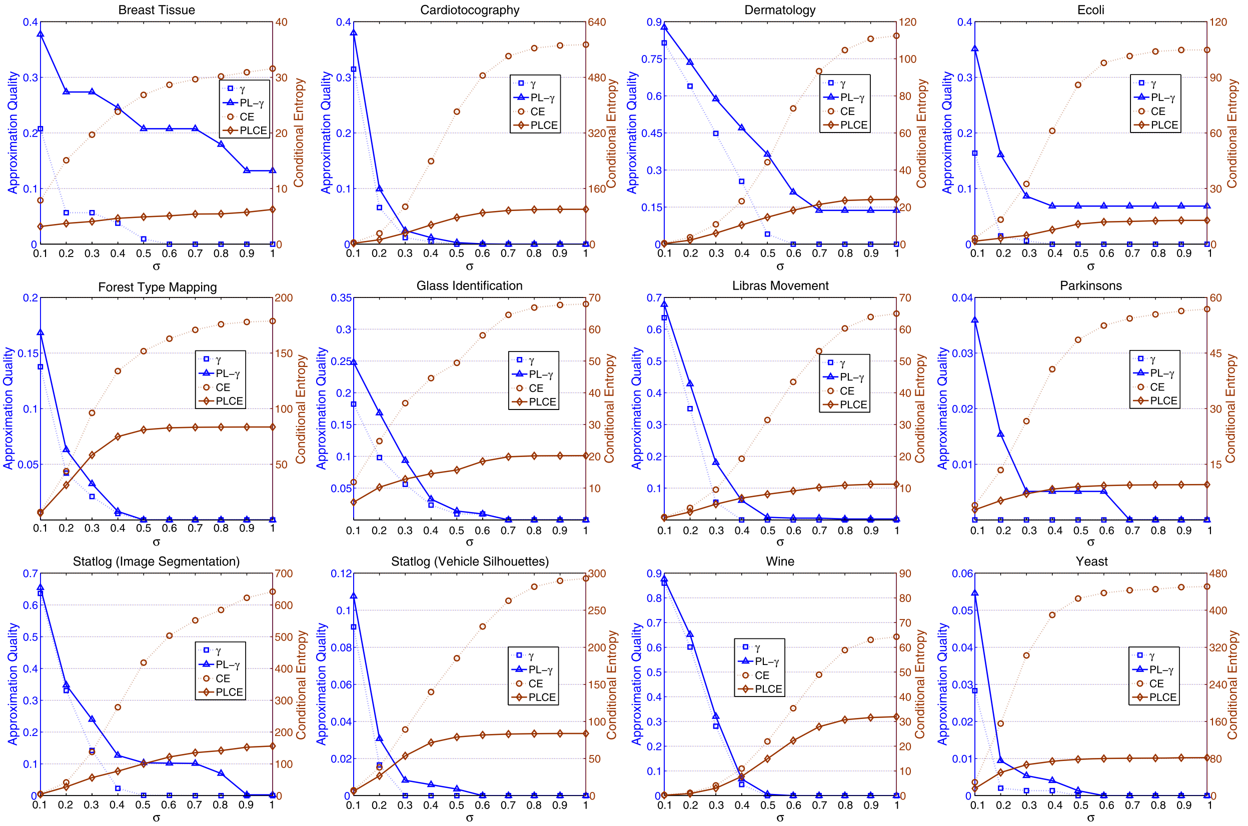

Fig. 3. Comparisons based on approximation qualities and conditional entropies.

1. If the value of σ increases, then the increasing trends have been obtained for both approximation quality and pseudo-label approximation quality, the decreasing trends have been obtained for both conditional entropy and pseudo-label conditional entropy. In other words, our pseudo-label strategy will not change the trends of the variations of approximation quality and conditional entropy.
2. For the 10 used values of σ, the values of pseudo-label approximation qualities are greater than or equal to those of traditional approximation qualities. Take for instance “Statlog (Image Segmentation)”, if $\sigma = 0.3$, then $\gamma_A(d) = 0.1420$ while $\gamma_A^{\mathrm{PL}}(d) = 0.2398$. Such observation implies that our pseudo-label strategy is useful in improving the degree of certain belongingness from the viewpoint of rough set. This observation is corresponding to what has been addressed in Proposition 4.
3. The pseudo-label conditional entropies are less than or equal to the traditional conditional entropies. Such case is significant when $\sigma \geq 0.3$ for all the data sets we tested. Take for instance “Ecoli”, if $\sigma = 0.3$, then $\mathrm{ENT}_A(d) = 32.4848$ and $\mathrm{ENT}_A^{\mathrm{PL}}(d) = 4.7554$. In other words, our pseudo-label strategy will effectively decrease the uncertainty degree from the viewpoint of neighborhood based information theory. This observation corresponds to what has been addressed in Proposition 5.
4. For each data set, the value of approximation quality achieves 0 when $\sigma = 1$. This is mainly because all samples in the data have been grouped into the neighborhood of each sample and then the lower approximation of each decision class may be an empty set. Correspondingly, no certainty is obtained. Nevertheless, the value of pseudo-label approximation quality may be greater than 0. For example, consider “Dermatology”, if $\sigma = 1$, then $\gamma_A(d) = 0$ and $\gamma_A^{\mathrm{PL}}(d) = 0.1366$. The reason is that though the samples cannot be distinguished by traditional neighborhood relation, the pseudo labels obtained from k-means clustering will provide us some degrees of discrimination for constructing pseudo-label neighborhood relation. Such extreme case indicates that pseudo-label strategy is effective in improving the certainty from the viewpoint of rough set.

In Fig. 4, “NDER” refers to the neighborhood decision error rate, “PLNDER” refers to the pseudo-label neighborhood decision error rate, “CDI” refers to the conditional discrimination index, and “PLCDI” refers to the pseudo-label conditional discrimination index.

With a deep investigation of Fig. 4, it is not difficult to observe the following.

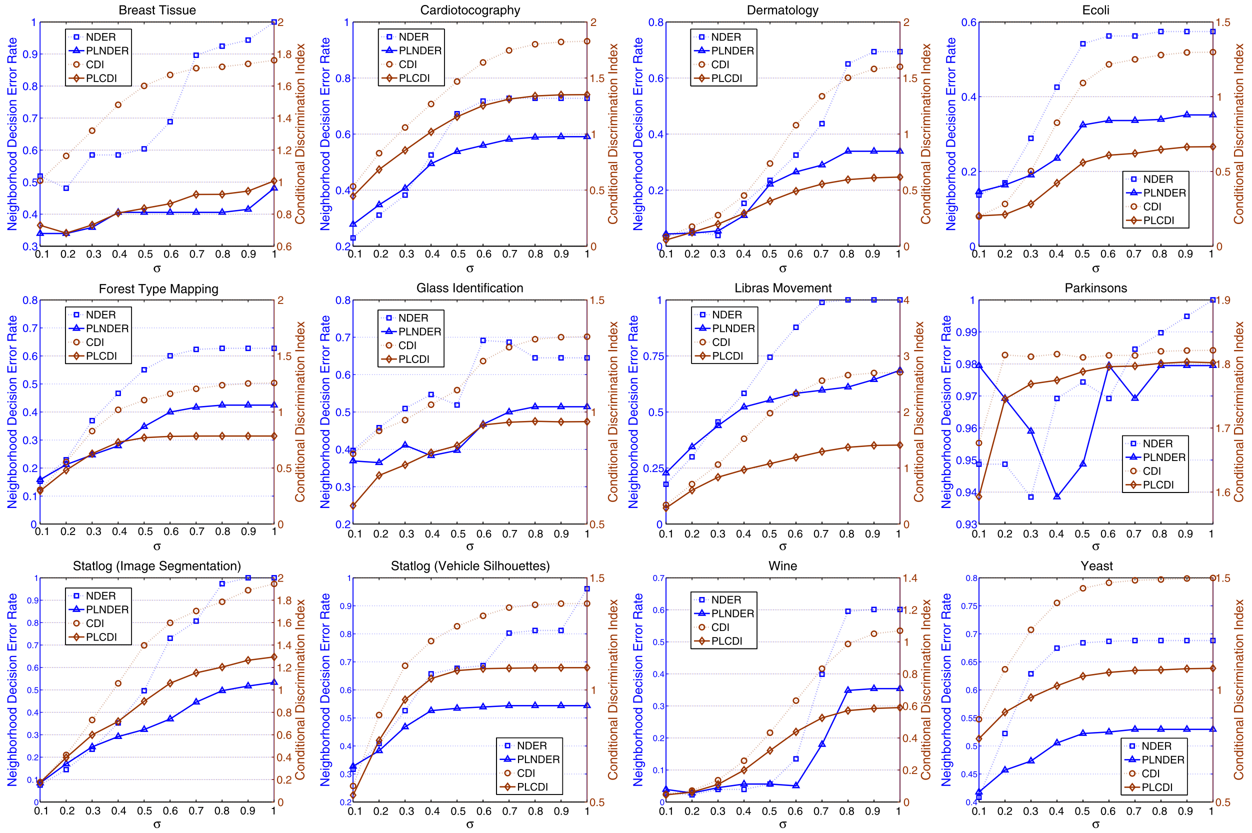

Fig. 4. Comparisons based on neighborhood decision error rates and conditional discrimination indexes.

1. If the value of σ increases, then both neighborhood decision error rate and pseudo-label neighborhood decision error rate show increasing trends, and both conditional discrimination index and pseudo-label conditional discrimination index show increasing trends. It should be noticed that such two trends are not necessarily monotonic. Take “Breast Tissue” as an example, if $\sigma = 0.1$, then $\mathrm{NDER}_A(d) = 0.5189$; if $\sigma = 0.2$, then $\mathrm{NDER}_A(d) = 0.4881$. Another example can be found in “Parkinsons”, if $\sigma = 0.2$, then $\mathrm{H}_A(d) = 1.8140$; if $\sigma = 0.3$, then $\mathrm{H}_A(d) = 1.8114$.
2. Generally speaking, the values of pseudo-label neighborhood decision error rates are less than those of neighborhood decision error rates. Moreover, with the greater value of σ, a significant difference between such two values is witnessed. Take for instance “Wine”, if $\sigma = 0.6$, then $\mathrm{NDER}_A(d) = 0.1348$ and $\mathrm{NDER}_A^{\mathrm{PL}}(d) = 0.0506$. Moreover, if $\sigma = 0.7$, then $\mathrm{NDER}_A(d) = 0.3989$ and $\mathrm{NDER}_A^{\mathrm{PL}}(d) = 0.1798$. Such case indicates that our pseudo-label strategy contributes a better classification performance if the value of σ is set to be higher. This is mainly because most of the samples whose labels are different from the true label of the test sample have be deleted by pseudo-label strategy.
3. The values of pseudo-label conditional discrimination indexes are less than or equal to those of conditional discrimination indexes. An example can be observed in “Parkinsons”, if $\sigma = 0.1$, then $\mathrm{H}_A(d) = 1.6766$ and $\mathrm{H}_A^{\mathrm{PL}}(d) = 1.5928$. That is to say, pseudo-label strategy is useful for decreasing the uncertainty in neighborhood decision system because pseudo-label strategy may offer a better neighborhood relation for distinguishing the samples. This observation corresponds to what has been addressed in Proposition 7.

In the following, the Wilcoxon signed rank test [8,60] will be selected for comparing the traditional neighborhood strategy and pseudo-label neighborhood strategy. Wilcoxon signed rank test is a non-parametric alternative to the paired t-test. From the viewpoint of statistical theory, this test is safer since it does not assume normal distributions. The purpose of our computation is trying to reject the null-hypothesis that traditional strategy and pseudo-label strategy perform equally well from the viewpoints of measures considered in this paper.

Take for instance “Breast Tissue”, the values of approximation qualities in terms of 10 values of σ are: 0.2075, 0.0566, 0.0566, 0.0377, 0.0094, 0.0000, 0.0000, 0.0000, and 0.0000; the values of pseudo-label approximation qualities in terms of 10 values of σ are: 0.3774, 0.2736, 0.2736, 0.2453, 0.2075, 0.2075, 0.2075, 0.1792, 0.1321, and 0.1321. Therefore, the corresponding p-value of Wilcoxon signed rank test is 0.0020. p-value is the probability of observing the given result, or one more extreme, by chance if the null hypothesis is true. The detailed results of p-values are shown in Table 4.

Table 4
p-value of Wilcoxon signed rank test based on measures (poorer p-values are italic).

ID	γ & PL-γ	CE & PLCE	CDI & PLCDI	NDER & PLNDER
1	0.0020	0.0020	0.0020	0.0020
2	0.0313	0.0020	0.0020	0.0469
3	0.0020	0.0020	0.0020	0.0215
4	0.0020	0.0020	0.0039	0.0059
5	*0.1250*	0.0020	0.0020	0.0039
6	*0.0625*	0.0020	0.0020	0.0020
7	0.0020	0.0020	0.0020	0.0195
8	0.0313	0.0020	0.0020	*0.6816*
9	0.0020	0.0020	0.0020	0.0273
10	*0.0625*	0.0020	0.0020	0.0039
11	*0.0625*	0.0020	0.0020	*0.1523*
12	*0.0625*	0.0020	0.0020	0.0039

Assume that if the significance level is set by 0.05, then we reject the null-hypothesis. Following the results of Table 4, we can observe: for the comparison between conditional entropy and pseudo-label conditional entropy, all obtained p-values are less than 0.05; for the comparison between conditional discrimination index and pseudo-label conditional discrimination index, all obtained p-values are also less than 0.05. These results indicate that traditional neighborhood approach and pseudo-label neighborhood approach do not perform equally well from the viewpoints of conditional entropy and conditional discrimination index. Moreover, for the comparison between approximation quality and pseudo-label approximation quality, 5 out of 12 p-values are greater than 0.05, while for the comparison between neighborhood decision error rate and pseudo-label neighborhood decision error rate, 2 out of 12 p-values are greater than 0.05, we can conclude that traditional neighborhood approach and pseudo-label neighborhood approach possibly do not perform equally well from the viewpoints of approximation quality and decision error rate.

The above analyses show that traditional neighborhood approach and pseudo-label neighborhood approach are so different in terms of the measures.

4.2.2. Comparisons of reducts

In this experiment, we will compare the reducts generated by traditional neighborhood approach and pseudo-label neighborhood approach, respectively.

To test the performances of reducts, 5-folder cross-validation is employed: in each iteration, 80% of the samples in data form the training set for computing reducts, and the rest of the 20% samples are considered as the test samples for evaluations, i.e., use attributes in reducts derived by training set to compute four measures over test samples. The above process is repeated 5 times. Then, the mean value of each measure is recorded. The final results are displayed in Figs. 5 and 6.

In Fig. 5, "γ-reduct" denotes the results based on approximation quality-reduct [12], "PL-γ-reduct" denotes the results based on pseudo-label approximation quality-reduct, "CE-reduct" denotes the results based on conditional entropy-reduct [6], and "PLCE-reduct" denotes the results based on pseudo-label conditional entropy-reduct.

With a careful investigation of Fig. 5, it is not difficult to observe the following.

1. If the value of σ increases, then both the values of approximation qualities derived from "γ-reduct" and pseudo-label approximation qualities derived from "PL-γ-reduct" are decreasing though such trend is not necessarily monotonic. Meanwhile, both the conditional entropies derived from "CE-reduct" and the pseudo-label conditional entropies derived from "PLCE-reduct" are increasing. Therefore, the reducts will not change the trends of the variations of approximation quality and conditional entropy for both traditional and pseudo-label approaches.
2. The pseudo-label approximation qualities derived from "PL-γ-reduct" are greater than approximation qualities derived from "γ-reduct". Let us take "Libras Movement" as an example, if $\sigma = 0.3$, then with the reducts, the approximation quality and pseudo-label approximation quality obtained over the test samples are 0.0556 and 0.2806, respectively. This is mainly because: before finding reducts, pseudo-label approximation qualities are greater than traditional approximation qualities (it has been shown in Fig. 3); the constraints of "γ-reduct" (see Definition 6) and "PL-γ-reduct" (see Definition 13) require that approximation qualities and pseudo-label approximation qualities will not decrease, at least.
3. Compared with "CE-reduct", attributes in "PLCE-reduct" will generate lower values of conditional entropies. For example, in "Statlog (Vehicle Silhouettes)", if $\sigma = 0.9$, then with reducts, the conditional entropy and pseudo-label conditional entropy over test samples are 55.3870 and 19.3757, respectively. The reason is similar to what has been analyzed in the case of approximation quality.

In Fig. 6, "NDER-reduct" refers to the neighborhood decision error rate-reduct [11], "PLNDER-reduct" refers to the pseudo-label neighborhood decision error rate-reduct, "CDI-reduct" refers to the conditional discrimination index-reduct [44], and "PLCDI-reduct" refers to the pseudo-label conditional discrimination index-reduct.

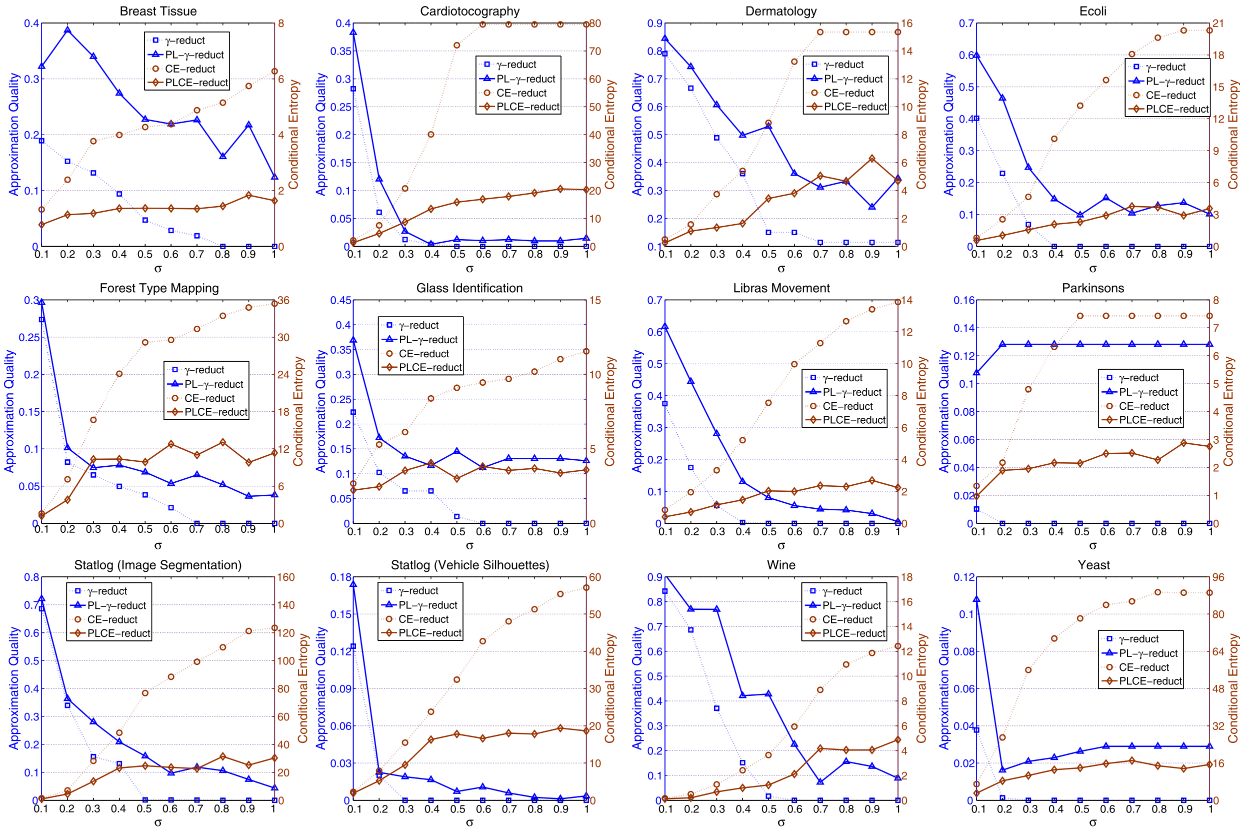

Fig. 5. Comparisons of reducts w.r.t. approximation quality and conditional entropy.

Since both "NDER-reduct" and "PLNDER-reduct" are designed for improving the classification performance, classification accuracy is then employed to evaluate the performances of them, respectively. Note that if "PLNDER-reduct" is used, then the classification accuracies are obtained by pseudo-label neighborhood classifier while if "NDER-reduct" is used, then the classification accuracies are obtained by traditional neighborhood classifier.

Based on Fig. 6, it is not difficult to observe the following.

1. With the increasing value of σ, both the conditional discrimination indexes obtained by "CDI-reduct" and the conditional discrimination indexes obtained by "PLCDI-reduct" increase though such trend is not necessarily monotonic. It indicates that reducts will not change the trend of variation of conditional discrimination index for both traditional and pseudo-label approaches. Meanwhile, the classification accuracies derived from attributes in "NDER-reduct" and "PLNDER-reduct" show decreasing trends.
2. For most of the data sets, if $\sigma < 0.4$, then there is no much difference between the classification accuracies derived from "PLNDER-reduct" and "NDER-reduct". However, if $\sigma \geq 0.4$, then the classification accuracies derived from "PLNDER-reduct" are significantly higher than those derived from "NDER-reduct". For example, in "Cardiotocography", if $\sigma = 0.6$, then the classification accuracies derived from "NDER-reduct" and "PLNDER-reduct" are 0.3170 and 0.4069, respectively. Such case shows that "PLNDER-reduct" is more effective than "NDER-reduct" if larger scale of σ is considered.
3. The pseudo-label conditional discrimination indexes derived from "PLCDI-reduct" are significantly lower than conditional discrimination indexes derived from "CDI-reduct". Take "Dermatology" as an example, if $\sigma = 0.5$, then with reducts, the values of conditional discrimination index and pseudo-label conditional discrimination index over test samples are 0.7008 and 0.4890, respectively.

Moreover, similar to Section 4.2.1, the Wilcoxon signed rank test [8] is also selected for comparing the reducts generated by traditional neighborhood and pseudo-label neighborhood approaches. This computation aims to reject the null-hypothesis that traditional neighborhood and pseudo-label neighborhood perform equally well for computing reducts. The detailed p-values are shown in Table 5.

Assume that if the significance level is given by 0.05, then we reject the null-hypothesis. For the reducts constrained by measures of approximation quality, conditional entropy and conditional discrimination index, all the p-values are less than 0.05 while for the measure of neighborhood decision error rate, 9 out of 12 p-values are less than 0.05. These results

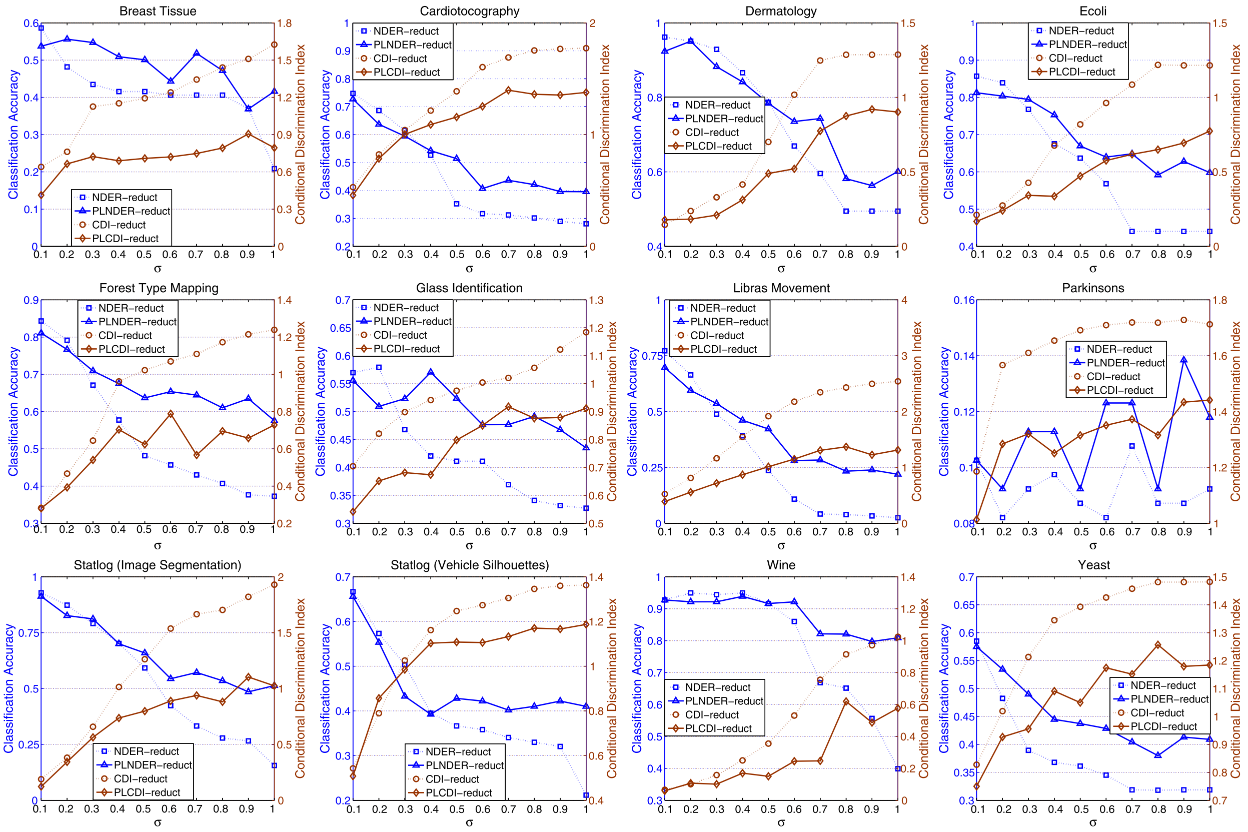

Fig. 6. Comparisons of reducts w.r.t. classification accuracy and conditional discrimination index.

Table 5
p-value of Wilcoxon signed rank test based on reducts (poorer p-values are italic).

ID	γ-reduct & PL-γ-reduct	CE-reduct & PLCE-reduct	CDI-reduct & PLCDI-reduct	NDER-reduct & PLNDER-reduct
1	0.0020	0.0020	0.0020	0.0137
2	0.0020	0.0020	0.0020	*0.0645*
3	0.0020	0.0020	0.0039	*0.2324*
4	0.0020	0.0020	0.0020	0.0371
5	0.0020	0.0020	0.0020	0.0098
6	0.0020	0.0020	0.0020	0.0195
7	0.0020	0.0020	0.0020	0.0293
8	0.0020	0.0020	0.0020	0.0039
9	0.0020	0.0020	0.0020	0.0371
10	0.0020	0.0020	0.0137	*0.1602*
11	0.0020	0.0020	0.0039	*0.1680*
12	0.0020	0.0020	0.0020	0.0039

indicate that traditional neighborhood approach and pseudo-label neighborhood approach do not perform equally well from the viewpoint of computing reducts.

5. Conclusions

In this paper, a pseudo-label strategy has been introduced into the neighborhood rough data analysis. Different from the traditional construction of neighborhood, our pseudo-label based neighborhood is obtained by not only the distance over condition attributes, but also the pseudo labels of samples generated by condition attributes. The pseudo-label approach may help us to improve the discrimination between samples. The experimental results demonstrate that pseudo-label strategy is useful in decreasing the uncertainties in neighborhood decision system. Furthermore, it is also shown that the reducts obtained by pseudo-label strategy are also superior to the reducts derived by traditional neighborhood way in terms of four different measures.

The following topics are challenges for our further research.

1. In this paper, the pseudo labels of samples are derived from only *k*-means clustering. More approaches to produce the pseudo-labels will be further employed in constructing pseudo-label neighborhood rough set.
2. Pseudo-label strategy can also be explored in other extended rough set models.
3. The pseudo-label neighborhood classifier will be compared with some other popular classifiers.

Acknowledgements

This work is supported by the Natural Science Foundation of China (Nos. 61572242, 61502211, 61503160), Open Project Foundation of Intelligent Information Processing Key Laboratory of Shanxi Province (No. 2014002), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX18_2333).

We would like to thank Eric Appiah Mantey and Selase Tawiah Kwawu for helping us improve the language quality of this paper.

References

[1] D.G. Chen, Y.Y. Yang, Attribute reduction for heterogeneous data based on the combination of classical and fuzzy rough set models, IEEE Trans. Fuzzy Syst. 22 (2014) 1325–1334.
[2] H.M. Chen, T.R. Li, Y. Cai, C. Luo, H. Fujita, Parallel attribute reduction in dominance-based neighborhood rough set, Inf. Sci. 373 (2016) 351–368.
[3] Y.H. Chen, Y.Y. Yao, Multiview intelligent data analysis based on granular computing, in: Proceedings of the 2006 IEEE International Conference on Granular Computing, Atlanta, 2006, pp. 281–286.
[4] J.H. Dai, S.C. Gao, G.J. Zheng, Generalized rough set models determined by multiple neighborhoods generated from a similarity relation, Soft Comput. 22 (2018) 2081–2094.
[5] T. Denoeux, S. Sriboonchitta, O. Kanjanatarakul, Evidential clustering of large dissimilarity data, Knowl.-Based Syst. 106 (2016) 179–195.
[6] J.H. Dai, W.T. Wang, H.W. Tian, L. Liu, Attribute selection based on a new conditional entropy for incomplete decision systems, Knowl.-Based Syst. 39 (2013) 207–213.
[7] J.H. Dai, Q. Xu, W.T. Wang, H.W. Tian, Conditional entropy for incomplete decision systems and its application in data mining, Int. J. Gen. Syst. 41 (2012) 713–728.
[8] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[9] H. Ge, L.S. Li, Y. Xu, C.J. Yang, Quick general reduction algorithms for inconsistent decision tables, Int. J. Approx. Reason. 82 (2017) 56–80.
[10] Q.H. Hu, J.F. Liu, D.R. Yu, Mixed feature selection based on granulation and approximation, Knowl.-Based Syst. 21 (2008) 294–304.
[11] Q.H. Hu, W. Pedrycz, D.R. Yu, J. Lang, Selecting discrete and continuous features based on neighborhood decision error minimization, IEEE Trans. Syst. Man Cybern., Part B 40 (2010) 137–150.
[12] Q.H. Hu, D.R. Yu, J.F. Liu, C.X. Wu, Neighborhood rough set based heterogeneous feature subset selection, Inf. Sci. 178 (2008) 3577–3594.
[13] Q.H. Hu, D.R. Yu, Z.X. Xie, Neighborhood classifiers, Expert Syst. Appl. 34 (2008) 866–876.
[14] Q.H. Hu, D.R. Yu, Z.X. Xie, J.F. Liu, Fuzzy probabilistic approximation spaces and their information measures, IEEE Trans. Fuzzy Syst. 16 (2006) 549–551.
[15] Q.H. Hu, L. Zhang, D.G. Chen, W. Pedrycz, D.R. Yu, Gaussian kernel based fuzzy rough sets: model, uncertainty measures and applications, Int. J. Approx. Reason. 51 (2010) 453–471.
[16] Q.H. Hu, L.J. Zhang, Y.C. Zhou, W. Pedrycz, Large-scale multi-modality attribute reduction with multi-kernel fuzzy rough sets, IEEE Trans. Fuzzy Syst. 26 (2018) 226–238.
[17] H.J. Jia, S.F. Ding, M. Heng, W.Q. Xing, Spectral clustering with neighborhood attribute reduction based on information entropy, J. Comput. 9 (2014) 1316–1324.
[18] X.Y. Jia, L. Shang, B. Zhou, Y.Y. Yao, Generalized attribute reduct in rough set theory, Knowl.-Based Syst. 91 (2016) 204–218.
[19] Y.G. Jing, T.R. Li, C. Luo, S.J. Horng, G.Y. Wang, Z. Yu, An incremental approach for attribute reduction based on knowledge granularity, Knowl.-Based Syst. 104 (2016) 24–38.
[20] Y.G. Jing, T.R. Li, J.F. Huang, Y.Y. Zhang, An incremental attribute reduction approach based on knowledge granularity under the attribute generalization, Int. J. Approx. Reason. 76 (2016) 80–95.
[21] H.R. Ju, H.X. Li, X.B. Yang, X.Z. Zhou, B. Huang, Cost-sensitive rough set: a multi-granulation approach, Knowl.-Based Syst. 123 (2017) 137–153.
[22] W.W. Li, Z.Q. Huang, X.Y. Jia, X.Y. Cai, Neighborhood based decision-theoretic rough set models, Int. J. Approx. Reason. 69 (2016) 1–17.
[23] J.Z. Li, X.B. Yang, X.N. Song, J.H. Li, P.X. Wang, D.J. Yu, Neighborhood attribute reduction: a multi-criterion approach, Int. J. Mach. Learn. Cybern. (2017), https://doi.org/10.1007/s13042-017-0758-5.
[24] J.Y. Liang, F. Wang, C.Y. Dang, Y.H. Qian, An efficient rough feature selection algorithm with a multi-granulation view, Int. J. Approx. Reason. 53 (2012) 912–926.
[25] G.P. Lin, J.Y. Liang, Y.H. Qian, Uncertainty measures for multigranulation approximation space, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 23 (2015) 443–457.
[26] Y.J. Lin, Q.H. Hu, J.H. Liu, J.K. Chen, J. Duan, Multi-label feature selection based on neighborhood mutual information, Appl. Soft Comput. 38 (2016) 244–256.
[27] Y.J. Lin, J.J. Li, P.R. Lin, G.P. Lin, J.K. Chen, Feature selection via neighborhood multi-granulation fusion, Knowl.-Based Syst. 67 (2014) 162–168.
[28] D. Liu, D.C. Liang, Incremental learning researches on rough set theory: status and future, Int. J. Rough Sets, Data Anal. 1 (2014) 99–112.
[29] X.M. Liu, D.M. Zhao, J.T. Zhou, W. Gao, H.F. Sun, Image interpolation via graph-based Bayesian label propagation, IEEE Trans. Image Process. 23 (2014) 1084–1096.
[30] Y. Liu, W.L. Huang, Y.L. Jiang, Z.Y. Zeng, Quick attribute reduct algorithm for neighborhood rough set model, Inf. Sci. 271 (2014) 65–81.
[31] F. Min, F.L. Liu, L.Y. Wen, Z.H. Zhang, Tri-partition cost-sensitive active learning through kNN, Soft Comput. 10 (2017) 1–16.
[32] J.M. Ma, Y.Y. Yao, Rough set approximations in multi-granulation fuzzy approximation spaces, Fundam. Inform. 142 (2015) 145–160.
[33] J. Qian, C.Y. Dang, X.D. Yue, N. Zhang, Attribute reduction for sequential three-way decisions under dynamic granulation, Int. J. Approx. Reason. 85 (2017) 196–216.
[34] Y.H. Qian, X.Y. Liang, G.P. Lin, Q. Guo, J.Y. Liang, Local multigranulation decision-theoretic rough sets, Int. J. Approx. Reason. 82 (2017) 119–137.
[35] Y.H. Qian, X.Y. Liang, Q. Wang, J.Y. Liang, B. Liu, A. Skowron, Y.Y. Yao, J.M. Ma, C.Y. Dang, Local rough set: a solution to rough data analysis in big data, Int. J. Approx. Reason. 97 (2018) 38–63.
[36] Y.H. Qian, J.Y. Liang, W. Pedrycz, C.Y. Dang, An efficient accelerator for attribute reduction from incomplete data in rough set framework, Pattern Recognit. 44 (2011) 1658–1670.
[37] Y.H. Qian, J.Y. Liang, W. Pedrycz, C.Y. Dang, Positive approximation: an accelerator for attribute reduction in rough set theory, Artif. Intell. 174 (2010) 597–618.

[38] Y.H. She, Y.Y. Yao, Rough set models in multigranulation spaces, Inf. Sci. 327 (2016) 40–56.
[39] W.H. Shu, W.B. Qian, An incremental approach to attribute reduction from dynamic incomplete decision systems in rough set theory, Data Knowl. Eng. 100 (2015) 116–132.
[40] R. Susmaga, R. Słowiński, Generation of rough sets reducts and constructs based on inter-class and intra-class information, Fuzzy Sets Syst. 274 (2015) 124–142.
[41] J.L. Tang, X. Hu, H.J. Gao, H. Li, Discriminant analysis for unsupervised feature selection, in: Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, 2014, pp. 938–946.
[42] J.L. Tang, H. Liu, Unsupervised feature selection for linked social media data, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, 2012, pp. 904–912.
[43] M. Wang, F. Min, Z.H. Zhang, Y.X. Wu, Active learning through density clustering, Expert Syst. Appl. 85 (2017) 305–317.
[44] C.Z. Wang, Q.H. Hu, X.Z. Wang, D.G. Chen, Y.H. Qian, Z. Dong, Feature selection based on neighborhood discrimination index, IEEE Trans. Neural Netw. Learn. Syst. 29 (2018) 2986–2999.
[45] C.Z. Wang, M.W. Shao, Q. He, Y.H. Qian, Y.L. Qi, Feature subset selection based on fuzzy neighborhood rough sets, Knowl.-Based Syst. 111 (2016) 173–179.
[46] F. Wang, J.Y. Liang, Y.H. Qian, Attribute reduction: a dimension incremental strategy, Knowl.-Based Syst. 39 (2013) 95–108.
[47] W. Wei, J.H. Wang, J.Y. Liang, X. Mi, C.Y. Dang, Compacted decision tables based attribute reduction, Knowl.-Based Syst. 86 (2015) 261–277.
[48] D.R. Wilson, T.R. Martinez, Improved heterogeneous distance functions, J. Artif. Intell. Res. 6 (1997) 1–34.
[49] K. Wu, K.H. Yap, Fuzzy SVM for content-based image retrieval: a pseudo-label support vector machine framework, IEEE Comput. Intell. Mag. 1 (2006) 10–16.
[50] X.J. Xie, X.L. Qin, A novel incremental attribute reduction approach for dynamic incomplete decision systems, Int. J. Approx. Reason. 93 (2018) 443–462.
[51] S.P. Xu, P.X. Wang, J.H. Li, X.B. Yang, X.J. Chen, Attribute reduction: an ensemble strategy, in: 2017 International Joint Conference on Rough Sets, Olsztyn, 2017, pp. 362–375.
[52] S.P. Xu, X.B. Yang, X.N. Song, H.L. Yu, Prediction of protein structural classes by decreasing nearest neighbor error rate, in: Proceedings of the 2015 International Conference on Machine Learning and Cybernetics, Guangzhou, 2015, pp. 7–13.
[53] S.P. Xu, X.B. Yang, E.C.C. Tsang, E.A. Mantey, Neighborhood collaborative classifiers, in: Proceedings of the 2016 International Conference on Machine Learning and Cybernetics, Jeju Island, 2016, pp. 470–476.
[54] S.P. Xu, X.B. Yang, H.L. Yu, D.J. Yu, J.Y. Yang, E.C.C. Tsang, Multi-label learning with label-specific feature reduction, Knowl.-Based Syst. 104 (2016) 52–61.
[55] X.B. Yang, Z.H. Chen, H.L. Dou, M. Zhang, J.Y. Yang, Neighborhood system based rough set: models and attribute reductions, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 20 (2012) 399–419.
[56] X.D. Yue, Y.F. Chen, D.Q. Miao, J. Qian, Tri-partition neighborhood covering reduction for robust classification, Int. J. Approx. Reason. 83 (2017) 371–384.
[57] X.B. Yang, Y. Qi, H.L. Yu, X.N. Song, J.Y. Yang, Updating multigranulation rough approximations with increasing of granular structures, Knowl.-Based Syst. 64 (2014) 59–69.
[58] X.B. Yang, Y.S. Qi, X.N. Song, J.Y. Yang, Test cost sensitive multigranulation rough set: model and minimal cost selection, Inf. Sci. 250 (2013) 184–199.
[59] X.B. Yang, S.P. Xu, H.L. Dou, X.N. Song, H.L. Yu, J.Y. Yang, Multigranulation rough set: a multiset based strategy, Int. J. Comput. Intell. Syst. 10 (2017) 277–292.
[60] X.B. Yang, Y.Y. Yao, Ensemble selector for attribute reduction, Appl. Soft Comput. 70 (2018) 1–11.
[61] D.J. Yu, J. Hu, X.W. Wu, H.B. Shen, J. Chen, Z.M. Tang, J. Yang, J.Y. Yang, Learning protein multi-view features in complex space, Amino Acids 44 (2013) 1365–1379.
[62] X.D. Yue, L.B. Cao, D.Q. Miao, Y.F. Chen, B. Xu, Multi-view attribute reduction model for traffic bottleneck analysis, Knowl.-Based Syst. 86 (2015) 1–10.
[63] A.P. Zeng, T.R. Li, D. Liu, J.B. Zhang, H.M. Chen, A fuzzy rough set approach for incremental feature selection on hybrid information systems, Fuzzy Sets Syst. 258 (2015) 39–60.
[64] W.R. Zeng, X.W. Chen, H. Cheng, Pseudo labels for imbalanced multi-label learning, in: Proceedings of the 2014 International Conference on Data Science and Advanced Analytics, Shanghai, 2014, pp. 25–31.
[65] J.B. Zhang, T.R. Li, D. Ruan, D. Liu, Neighborhood rough sets for dynamic data mining, Int. J. Intell. Syst. 27 (2012) 317–342.
[66] X. Zhang, C.L. Mei, D.G. Chen, J.H. Li, Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy, Pattern Recognit. 56 (2016) 1–15.
[67] H. Zhao, F. Min, W. Zhu, Test-cost-sensitive attribute reduction based on neighborhood rough set, in: Proceedings of the 2011 IEEE International Conference on Granular Computing, Kaohsiung, 2011, pp. 802–806.
[68] T.T. Zheng, L.Y. Zhu, Uncertainty measures of neighborhood system-based rough sets, Knowl.-Based Syst. 86 (2015) 57–65.
[69] P. Zhu, Q.Y. Wen, Information-theoretic measures associated with rough set approximations, Inf. Sci. 212 (2012) 33–43.
[70] P.F. Zhu, W.C. Zhu, Q.H. Hu, C.Q. Zhang, W.M. Zuo, Subspace clustering guided unsupervised feature selection, Pattern Recognit. 66 (2017) 364–374.