Research Article

Pseudolabel Decision-Theoretic Rough Set

Xun Wang,[1] Wendong Zhang,[1] Dun Liu,[2] Hualong Yu [iD],[1] Xibei Yang [iD],[1] and Xin Yang[3]

[1]*School of Computer, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu, 212003, China*
[2]*School of Economics and Management, Southwest Jiao Tong University, Chengdu, 610031, China*
[3]*School of Information Science and Technology, Southwest Jiao Tong University, Chengdu, 611756, China*

Correspondence should be addressed to Xibei Yang; zhenjiangyangxibei@163.com

In decision-theoretic rough set (DTRS), the decision costs are used to generate the thresholds for characterizing the probabilistic approximations. Similar to other rough sets, many generalized DTRS can also be formed by using different binary relations. Nevertheless, it should be noticed that most of the processes for calculating binary relations do not take the labels of samples into account, which may lead to the lower discrimination; for example, samples with different labels are regarded as indistinguishable. To fill such gap, the main contribution of this paper is to propose a pseudolabel strategy for constructing new DTRS. Firstly, a pseudolabel neighborhood relation is presented, which can differentiate samples by not only the neighborhood technique but also the pseudolabels of samples. Immediately, the pseudolabel neighborhood decision-theoretic rough set (PLNDTRS) can be constructed. Secondly, the problem of attribute reduction is explored, which aims to further reduce the PLNDTRS related decision costs. A heuristic algorithm is also designed to find such reduct. Finally, the clustering technique is employed to generate the pseudolabels of samples; the experimental results over 15 UCI data sets tell us that PLNDTRS is superior to DTRS without using pseudolabels because the former can generate lower decision costs. Moreover, the proposed heuristic algorithm is also effective in providing satisfied reducts. This study suggests new trends concerning cost sensitivity problem in rough data analysis.

1. Introduction

Different from classical rough set and its various generalizations [1–3], decision-theoretic rough set (DTRS) [4–6] has been demonstrated to be useful in many cost related problems [7–11]. Such model introduces not only the Bayesian decision but also the minimal risk into the construction of lower and upper approximations. Immediately, the thresholds for defining approximations have clear explanations.

From the viewpoint of granular computing [12], binary relation provides us with an effective mechanism for realizing information granulation. Therefore, the model of DTRS is used to characterize the relationship between the result of information granulation (information granules) and decision classes through considering the problem of costs. Consequently, DTRS may contain two important aspects: (1) information granulation and (2) decision costs. On the one hand, different types of information granulations may be used to construct different models [4, 13–19]. For example, the classical DTRS is formed based on the indiscernibility relation [4]; similar to Pawlak's rough set, such model is only useful in analyzing categorical data; Li et al. [16] have proposed a neighborhood based DTRS, in which the result of information granulation is expressed by neighborhoods of samples; Song et al. [15] have proposed a fuzzy based DTRS, in which the result of information granulation is reflected by fuzzy relation. These representative results are suitable for analysing data with continuous or even mixed values. On the other hand, different problems related to costs also provide us with more directions for studying DTRS [20, 21]. For instance, Dou et al. [10] have taken the characteristics of multiplicity and variability of costs into consideration; they proposed three different multicost based DTRS; Liang et al. [22] have introduced the interval-value based cost into DTRS, which enlarges the application range of DTRS; to solve the problem of multiclass decision effectively, Yang et al. [23] have proposed a sequential three-way approach.

From discussions above, it should be emphasized that though different binary relations have been used in constructing DTRS, few of them take the labels of samples into account.
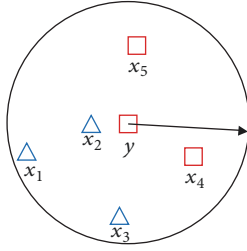
Figure 1: A small example.

Therefore, the ability to discriminate may not perform well. This is mainly because for any two samples their similarity is only determined by attributes or features regardless of the labels of such two samples. Consequently, samples with different labels may meet the requirement of binary relation, which will produce the extra decision costs [5].

What is more, in the studies of neighborhood systems [24–26], is that using distance functions is a very common and useful method. For example, Hu et al. [27] proposed a neighborhood based rough set model, which is easy to understand and implement; Li et al. [16] have replaced the equivalence class used in traditional DTRS by neighborhood relation. Correspondingly, we can use the neighborhood relation in this paper.

Take a binary classification task as an example (see Figure 1). There are 5 samples in the neighborhood of sample y except y itself. That is, any one of the samples x_i ($i = 1, 2, 3, 4, 5$) and y satisfies the neighborhood relation. In addition, samples x_1, x_2, and x_3 have the label of triangle, while x_4 and x_5 have the label of square. Obviously, y will be misclassified if the majority rule is employed; such misclassification may produce additional cost if the cost sensitivity problem is considered. To reduce this type of the misclassification and the corresponding cost, a natural thinking is to use a smaller scale of neighborhood [28]. However, it does not always work. For example, in Figure 1, if the value of radius keeps reducing, then it is possible that only x_2 is in the neighborhood of y if the reflexivity is ignored. In this case, y will also be misclassified.

Why reducing the value of radius is not always an effective way? This is mainly because the distance between samples is only determined by the condition attributes or features, while ignoring the label information of samples. The lower quality of attributes may not be good enough for providing better discrimination. From this point of view, new technique for measuring the similarity between samples, which considers the information provided by the labels, has become a necessity.

As we all know, in rough set theory, decision attribute offers us the labels of samples. However, these labels cannot be directly used, mainly because the labels over decision attribute are used to generate decision classes for constructing approximations. It is unreasonable to approximately characterize the decision classes by using the information which is derived from the decision attribute itself. For such reason, new sources of the labels of samples should be considered.

Fortunately, motivated by the results of pseudolabel strategy in unsupervised and semisupervised learning tasks [29–34], we know that the pseudolabels of samples provide us with another representation of labels. Therefore, it may be a useful attempt for us to introduce the pseudolabel strategy into DTRS.

Figure 2 reports the framework of our study. It must be emphasized that our research is based on neighborhood related DTRS. The reason is that neighborhood is suitable for dealing with complex data with simple structure. Furthermore, neighborhood offers us a natural structure of multi-granularity with the variation of radii. Obviously, following Figure 2, our main contribution is the pseudolabel strategy. In other words, the neighborhood of sample is determined by not only the binary relation, but also the pseudolabels. Correspondingly, the superiority of the pseudolabel strategy is that it can take the labels of samples into account, which the discrimination of samples can be improved.

The remainder of this paper is organized as follows. The basic knowledge about decision-theoretic rough set and neighborhood rough set is presented in Section 2. In Section 3, the model of pseudolabel neighborhood decision-theoretic rough set (PLNDTRS) is proposed. To further reduce the decision costs, the problem of attribute reduction is explored in Section 4. In Section 5, not only the experimental comparisons will be presented, but also the effectiveness of our strategy will be analyzed. The paper is ended with conclusions and perspectives for future work in Section 6.

2. Preliminary Knowledge

2.1. Decision-Theoretic Rough Set. In this section, some basic notions about DTRS [5, 16] are presented. Similar to the introduction of other rough sets, the concept of decision system should be firstly described for simplifying the subsequent discussions.

Definition 1. A decision system can be represented as $DS = < U, AT, d >$, in which $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of samples, called the universe; AT is the set of condition attributes; d is the decision attribute. $\forall x \in U$, $d(x)$ expresses the label of sample x, and $a(x)$ denotes the value that x holds on condition attribute a ($\forall a \in AT$).

Given a decision system DS, the classification task will be considered in this paper and then labels of samples are discrete. Consequently, an equivalence relation over d can be obtained such that $\text{IND}(d) = \{(x, y) \in U \times U : d(x) = d(y)\}$. Immediately, $U/\text{IND}(d) = \{X_1, X_2, \ldots, X_q\}$ is regarded as the partition determined by $\text{IND}(d)$. $\forall X_k \in U/\text{IND}(d)$, X_k is the set of samples with same label and then it is referred to as the k-th decision class. Specially, the decision class which contains sample x is denoted by $[x]_d$.

To know what is DTRS, the following notions should be given. Let $\Omega = \{X, \sim X\}$ indicate that a sample x is in X or out of X, that is, the states of x. The conditional probability of x in X can be obtained such that $\text{Pro}(X|[x]) = |X \cap [x]|/|[x]|$, in which $[x]$ is the information granule [35, 36] of x generated by using condition attributes and $|X|$ is the cardinal number
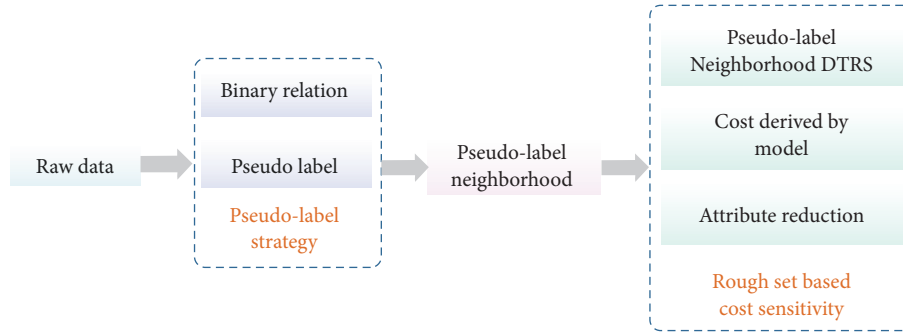
FIGURE 2: The framework of this paper.

TABLE 1: The loss function regarding the costs of three actions in two states.

	X	$\sim X$
a_P	λ_{PP}	λ_{PN}
a_B	λ_{BP}	λ_{BN}
a_N	λ_{NP}	λ_{NN}

of set X. For example, the equivalence classes derived by the equivalence relation over condition attributes AT can be regarded as information granules.

Furthermore, $\forall X \subseteq U$, three actions in $\mathscr{A} = \{a_P, a_B, a_N\}$ indicate x belongs to X or x belongs to X possibly or x does not belong to X. Therefore, the loss function regarding the costs of three actions in two different states is given in Table 1. In Table 1, λ_{*P} ($* \in \{P, B, N\}$) denote the costs for taking action a_P, a_B, or a_N when the x actual belongs to X, and λ_{*N} ($* \in \{P, B, N\}$) denote the costs for taking action a_P, a_B, or a_N when the x actual does not belong to X.

Following the costs shown in Table 1 and the conditional probability, the Bayesian costs for three different decisions can be defined as

$$\mathscr{R}_P = R\left(a_P \mid [x]\right)$$
$$= \lambda_{PP} \cdot \mathrm{Pro}\left(X \mid [x]\right) + \lambda_{PN} \cdot \mathrm{Pro}\left(\sim X \mid [x]\right);$$

$$\mathscr{R}_B = R\left(a_B \mid [x]\right)$$
$$= \lambda_{BP} \cdot \mathrm{Pro}\left(X \mid [x]\right) + \lambda_{BN} \cdot \mathrm{Pro}\left(\sim X \mid [x]\right);$$

$$\mathscr{R}_N = R\left(a_N \mid [x]\right)$$
$$= \lambda_{NP} \cdot \mathrm{Pro}\left(X \mid [x]\right) + \lambda_{NN} \cdot \mathrm{Pro}\left(\sim X \mid [x]\right).$$

(1)

Correspondingly, the following minimum-cost decision rules can be derived:

(P) If $\mathscr{R}_P \leq \mathscr{R}_B$ and $\mathscr{R}_P \leq \mathscr{R}_N$, then $x \in \mathrm{POS}(X)$;

(B) If $\mathscr{R}_B \leq \mathscr{R}_P$ and $\mathscr{R}_B \leq \mathscr{R}_N$, then $x \in \mathrm{BND}(X)$;

(N) If $\mathscr{R}_N \leq \mathscr{R}_P$ and $\mathscr{R}_N \leq \mathscr{R}_B$, then $x \in \mathrm{NEG}(X)$;

where $\mathrm{POS}(X)$ denotes the positive region of X, $\mathrm{BND}(X)$ denotes the boundary region of X, and $\mathrm{NEG}(X)$ denotes

the negative region of X. Samples in $\mathrm{POS}(X)$ indicate that these samples belong to X determinately, samples in $\mathrm{BND}(X)$ indicate that these samples belong to X possibly, and samples in $\mathrm{NEG}(X)$ indicate that these samples do not belong to X.

Since $\mathrm{Pro}(X \mid [x]) + \mathrm{Pro}(\sim X \mid [x]) = 1$, we also assume a reasonable loss function with the conditions such that $0 \leq \lambda_{PP} \leq \lambda_{BP} \leq \lambda_{NP}$ and $0 \leq \lambda_{NN} \leq \lambda_{BN} \leq \lambda_{PN}$; decision rules (P), (B), and (N) can be expressed as [10, 37]:

(P) If $\mathrm{Pro}(X \mid [x]) \geq \alpha$ and $\mathrm{Pro}(X \mid [x]) \geq \gamma$, then $x \in \mathrm{POS}(X)$;

(B) If $\mathrm{Pro}(X \mid [x]) \leq \alpha$ and $\mathrm{Pro}(X \mid [x]) \geq \beta$, then $x \in \mathrm{BND}(X)$;

(N) If $\mathrm{Pro}(X \mid [x]) \leq \beta$ and $\mathrm{Pro}(X \mid [x]) \leq \gamma$, then $x \in \mathrm{NEG}(X)$;

where three parameters α, β, and γ are

$$\alpha = \frac{\lambda_{PN} - \lambda_{BN}}{\left(\lambda_{PN} - \lambda_{BN}\right) + \left(\lambda_{BP} - \lambda_{PP}\right)};$$

$$\beta = \frac{\lambda_{BN} - \lambda_{NN}}{\left(\lambda_{BN} - \lambda_{NN}\right) + \left(\lambda_{NP} - \lambda_{BP}\right)};$$

$$\gamma = \frac{\lambda_{PN} - \lambda_{NN}}{\left(\lambda_{PN} - \lambda_{NN}\right) + \left(\lambda_{NP} - \lambda_{PP}\right)}.$$

(2)

In addition, if we assume the following condition for the loss function [5, 37]:

$$\frac{\left(\lambda_{NP} - \lambda_{BP}\right)}{\left(\lambda_{BN} - \lambda_{NN}\right)} > \frac{\left(\lambda_{BP} - \lambda_{PP}\right)}{\left(\lambda_{PN} - \lambda_{BN}\right)},$$

(3)

then we have $0 \leq \beta < \gamma < \alpha \leq 1$. In this case, the final decision rules associated with the conditional probability and thresholds can be obtained:

(P) If $\mathrm{Pro}(X \mid [x]) \geq \alpha$, then $x \in \mathrm{POS}(X)$;

(B) If $\beta < \mathrm{Pro}(X \mid [x]) < \alpha$, then $x \in \mathrm{BND}(X)$;

(N) If $\mathrm{Pro}(X \mid [x]) \leq \beta$, then $x \in \mathrm{NEG}(X)$.

If we denote $p = \text{Pro}(X \mid [x])$, then we have the following Bayesian costs for three different rules:

$$\text{Cost of positive rule} : p \cdot \lambda_{PP} + (1 - p) \cdot \lambda_{PN};$$

$$\text{Cost of boundary rule} : p \cdot \lambda_{BP} + (1 - p) \cdot \lambda_{BN}; \quad (4)$$

$$\text{Cost of negative rule} : p \cdot \lambda_{NP} + (1 - p) \cdot \lambda_{NN}.$$

From discussions above, the decision-theoretic based lower and upper approximations of $X \subseteq U$ are

$$\underline{X}^{(\alpha,\beta)} = \{x \in U : \text{Pro}(X \mid [x]) \geq \alpha\}; \quad (5)$$

$$\overline{X}^{(\alpha,\beta)} = \{x \in U : \text{Pro}(X \mid [x]) > \beta\}. \quad (6)$$

The pair $[\underline{X}^{(\alpha,\beta)}, \overline{X}^{(\alpha,\beta)}]$ is referred to as a decision-theoretic rough set of X, $\text{POS}(X) = \underline{X}^{(\alpha,\beta)}$ is the decision-theoretic positive region of X, $\text{BND}(X) = \overline{X}^{(\alpha,\beta)} - \underline{X}^{(\alpha,\beta)}$ is the decision-theoretic boundary region of X, and $\text{NEG}(X) = U - \overline{X}^{(\alpha,\beta)}$ is the decision-theoretic negative region of X.

2.2. Neighborhood Rough Set.

Neighborhood rough set is another generalization of classical rough set. There are two main advantages for neighborhood rough set: (1) it is suitable for dealing with continuous data or even mixed data, mainly because the neighborhood relation is constructed based on the consideration of distance between samples; (2) the scale of neighborhood provides us with a flexible technique to measure the granularity [38] and then the structure of multigranularity [39, 40] can be naturally formed.

Definition 2. Given a decision system DS and a neighborhood radius $\delta \in (0, 1]$, $\forall A \subseteq AT$, the neighborhood relation in terms of A can be defined as [27]:

$$\delta_A = \{(x, y) \in U \times U : \Delta_A(x, y) \leq \delta\}, \quad (7)$$

$\forall x \in U$; the neighborhood of x is then defined as

$$\delta_A(x) = \{y \in U : \Delta_A(x, y) \leq \delta\}, \quad (8)$$

where Δ is a metric function and the Euclidean distance is a widely accepted form such that

$$\Delta_A(x, y) = \sqrt{\sum_{a \in A}(a(x) - a(y))^2}. \quad (9)$$

Obviously, the neighborhood relation shown in Equation (7) is one of the binary relations; it is reflexive and symmetric. Moreover, it is not difficult to observe that the neighborhood relation produces neighborhood based information granule [41], in which samples are considered similar to the central sample. The size of the neighborhood is determined by the radius δ. If different values of δ are used, then different results of neighborhood relations will be generated. Generally speaking, smaller value of δ will generate finer neighborhood relation while greater value of δ will generate coarser neighborhood relation.

Definition 3. Given a decision system DS, $\forall X \subseteq U$, and $\forall A \subseteq AT$, the neighborhood lower and upper approximations of X with respect to A are defined as

$$\underline{N}_A(X) = \{x \in U : \delta_A(x) \subseteq X\}; \quad (10)$$

$$\overline{N}_A(X) = \{x \in U : \delta_A(x) \cap X \neq \emptyset\}. \quad (11)$$

The pair $[\underline{N}_A(X), \overline{N}_A(X)]$ is referred to as a neighborhood rough set of X.

Proposition 4. *[27] Given a decision system DS and two radii such that $\delta_1 \leq \delta_2$, then we have $\delta_1(x) \subseteq \delta_2(x)$ for each $x \in U$.*

Proposition 5. *[27] Given a decision system DS, $\forall A \subseteq AT$, $\forall x \in U$, and $\delta \in (0, 1]$, then we have $\delta_{AT}(x) \subseteq \delta_A(x)$.*

Proposition 6. *[27] Given a decision system DS, $\forall A \subseteq AT$, $\forall x \in U$, and $\delta \in (0, 1]$, then we have $\underline{N}_A(X) \subseteq \underline{N}_{AT}(X)$ and $\overline{N}_A(X) \supseteq \overline{N}_{AT}(X)$.*

Propositions 5 and 6 tell us that if the number of used attributes is increasing, then the size of the neighborhood will be reduced; it follows that the lower approximation will be expanded and the upper approximation will be narrowed.

2.3. Neighborhood Decision-Theoretic Rough Set.

One of the main motivations of neighborhood decision-theoretic rough set (NDTRS) is that the classical DTRS can only be used to deal with categorical data. Therefore, to further expand the applications of DTRS, the information granule $[x]$ in DTRS is replaced by neighborhood in NDTRS; that is, $\forall A \subseteq AT$; the conditional probability is $\text{Pro}(X \mid \delta_A(x)) = |X \cap \delta_A(x)|/|\delta_A(x)|$. Immediately, given a decision system DS, $\forall X \subseteq U$, and $\forall A \subseteq AT$, the neighborhood decision-theoretic based lower and upper approximations of X are [16]

$$\underline{N}_A^{(\alpha,\beta)}(X) = \{x \in U : \text{Pro}(X \mid \delta_A(x) \geq \alpha\}; \quad (12)$$

$$\overline{N}_A^{(\alpha,\beta)}(X) = \{x \in U : \text{Pro}(X \mid \delta_A(x) > \beta\}. \quad (13)$$

Correspondingly, the positive, boundary and negative regions of X are

$$\text{POS}_A^{(\alpha,\beta)}(X) = \underline{N}_A^{(\alpha,\beta)}(X);$$

$$\text{BND}_A^{(\alpha,\beta)}(X) = \overline{N}_A^{(\alpha,\beta)}(X) - \underline{N}_A^{(\alpha,\beta)}(X); \quad (14)$$

$$\text{NEG}_A^{(\alpha,\beta)}(X) = U - \overline{N}_A^{(\alpha,\beta)}(X).$$

3. Pseudolabel Neighborhood Decision-Theoretic Rough Set (PLNDTRS)

Though NDTRS uses the neighborhood relation to replace the indiscernibility relation for realizing information granulation, the derived neighborhoods are still closely related to the distances which are determined by the values of condition attributes or features. However, it is well known that the Euclidean distance does not always perform well for

distinguishing samples, especially in high-dimensional data. Therefore, new techniques have become the necessary issues to be addressed.

Following the construction of distance based neighborhood relation, that is, Equation (7), it is not difficult to observe that such strategy does not take the labels of samples into consideration. Therefore, it is possible that different samples with different labels may fall into the same neighborhood. From this point of view, if the labels of samples are used, then such situation may be relieved. Nevertheless, the raw labels of samples in decision system are not suitable for directly using, mainly because we cannot approximate the decision classes generated by decision attribute through using the label provided by the decision attribute itself. For such reason, we will further expand decision system DS by additional information of pseudolabels. The sources of pseudolabels may be derived from clustering analysis, classification analysis, label propagation [42], and so on.

Formally, a pseudolabel decision system can be represented as $DS^{PL} =< U, AT, d, d_{AT}^{PL} >$, in which U, AT, and d are similar to those in decision system DS, respectively; d_{AT}^{PL} is the pseudolabel decision attribute, $\forall x \in U$, $d_{AT}^{PL}(x)$ is the pseudolabel of x that is derived from a learning approach over AT. And the learning approach is k-means clustering in this paper.

Since both condition attributes and pseudolabels of samples exist in the pseudolabel neighborhood decision system, we can define the following pseudolabel neighborhood relation for replacing the traditional neighborhood relation shown in Equation (7).

Definition 7. Given a pseudolabel decision system DS^{PL} and a neighborhood radius $\delta \in (0, 1]$, $\forall A \subseteq AT$, the pseudolabel neighborhood relation is defined as

$$\delta_A^{PL} = \left\{(x, y) \in U \times U : \Delta_A(x, y) \le \delta \wedge d_A^{PL}(x) = d_A^{PL}(y)\right\}. \tag{15}$$

Following Definition 7, $\forall x \in U$, the pseudolabel neighborhood of x is then defined as

$$\delta_A^{PL}(x) = \left\{y \in U : \Delta_A(x, y) \le \delta \wedge d_A^{PL}(x) = d_A^{PL}(y)\right\}. \tag{16}$$

Obviously, in Definition 7, two constraints have been employed to construct the pseudolabel neighborhood relation: (1) similar to previous neighborhood relation shown in Equation (7), the distance between two samples should be less than or equal to the given radius; (2) two samples must have same pseudolabel.

Proposition 8. *Given a pseudolabel decision system DS^{PL}, $\forall A \subseteq AT$, we have $\delta_A^{PL}(x) \subseteq \delta_A(x)$.*

Proof. It can be derived directly from Equations (8) and (16). □

The above proposition tells us that, by comparing the traditional approach, our pseudolabel strategy can derive a finer

neighborhood relation which provides higher performance of discrimination, mainly because the samples with different pseudolabels should be deleted from the neighborhood.

Naturally, following Equation (16), the pseudolabel neighborhood decision-theoretic rough set (PLNDTRS) can be formed as Definition 9 shows.

Definition 9. Given a pseudolabel decision system DS^{PL}, $\forall X \subseteq U$, and $\forall A \subseteq AT$, the pseudolabel neighborhood lower and upper approximations of X are

$$\underline{PLN}_A^{(\alpha,\beta)}(X) = \left\{x \in U : \mathrm{Pro}\left(X \mid \delta_A^{PL}(x)\right) \ge \alpha\right\}; \tag{17}$$

$$\overline{PLN}_A^{(\alpha,\beta)}(X) = \left\{x \in U : \mathrm{Pro}\left(X \mid \delta_A^{PL}(x)\right) > \beta\right\}. \tag{18}$$

The pair $[\underline{PLN}_A^{(\alpha,\beta)}(X), \overline{PLN}_A^{(\alpha,\beta)}(X)]$ is referred to as a PLNDTRS of X. Correspondingly, the pseudolabel positive, boundary, and negative regions of X are

$$\mathrm{PLPOS}_A^{(\alpha,\beta)}(X) = \underline{PLN}_A^{(\alpha,\beta)}(X);$$

$$\mathrm{PLBND}_A^{(\alpha,\beta)}(X) = \overline{PLN}_A^{(\alpha,\beta)}(X) - \underline{PLN}_A^{(\alpha,\beta)}(X); \tag{19}$$

$$\mathrm{PLNEG}_A^{(\alpha,\beta)}(X) = U - \overline{PLN}_A^{(\alpha,\beta)}(X).$$

Example 10. For readers' convenience, we can take an example to show the difference between NDTRS and PLNDTRS. Figure 3 shows us a binary classification problem. Two classes of samples are represented by blue "·" and red "+", respectively.

(1) Let us investigate subfigure (a); suppose that x is a testing sample for classification; based on the given radius δ, 8 samples fall into the neighborhood of x except x itself; they are $y1, \ldots, y_8$. Obviously, samples y_1, y_4, and y_6 have the label of blue "·" and samples y_2, y_3, y_5, y_7, and y_8 have the label of red "+". We then conclude that x will be misclassified if the majority rule is employed, mainly because the majority rule classifies sample x into the class of red "+" while the real label of x is blue "·".

Furthermore, if the value of α is 0.3, then by Euation (12), we can conclude that x belongs to the lower approximations of both two different classes. The reason is that $\mathrm{Pro}(X_1 \mid \delta_A(x)) = 0.44$ and $\mathrm{Pro}(X_2 \mid \delta_A(x)) = 0.56$ if X_1 indicates the class of blue "·" while X_2 indicates the red "+". Obviously, it is an inconsistent case and there is a conflict between two different lower approximations.

(2) Let us investigate subfigure (b) which shows us both the real labels and the pseudolabels of samples. The real labels are same to those in subfigure (a); the pseudo labels of samples are denoted by "△" and "○", respectively. Obviously, y_1, y_2, y_3, y_7, and y_8 have the same pseudolabel ("○") while y_4, y_5, y_6, and x have another same pseudolabel ("△"). Based on the expression of Definition 7, y_1, y_2, y_3, y_7, and y_8 should be deleted from the neighborhood of x, mainly because, by comparing x, these samples have different pseudolabels. Immediately, y_4, y_5, and y_6 are in the pseudolabel neighborhood of sample x. Therefore, sample x can be correctly classified if the majority rule is used.
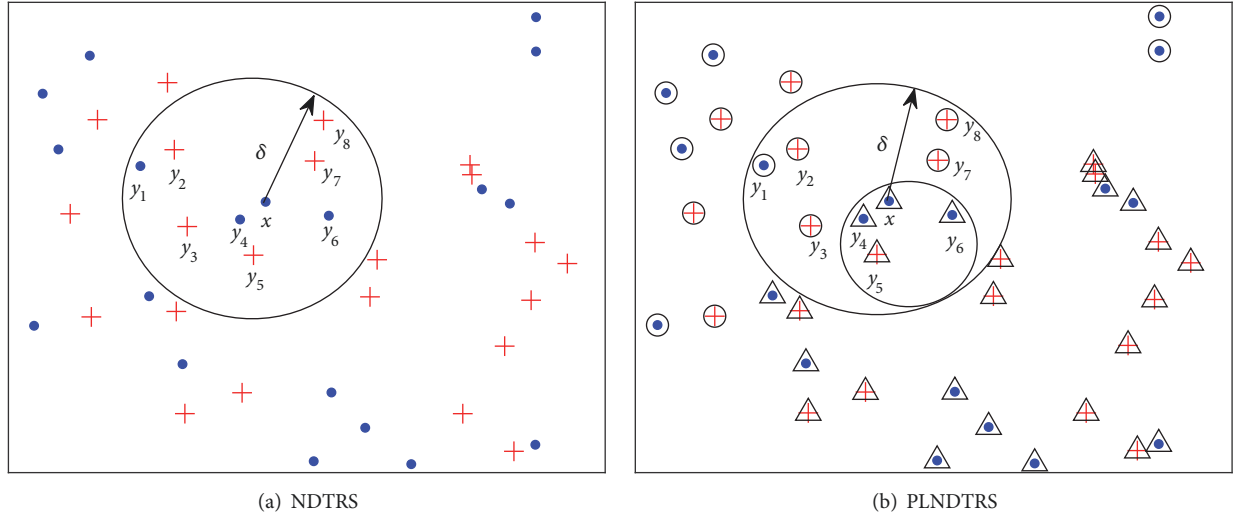
(a) NDTRS

(b) PLNDTRS

FIGURE 3: The difference between NDTRS and PLNDTRS.

In addition, if the value of α is 0.3, then, by Equation (17), we can observe that x belongs to the lower approximation of the class of blue "." while it does not belong to the lower approximation of the class of red "+", mainly because $\text{Pro}(X_1 \mid \delta_A^{PL}(x)) = 0.75$ and $\text{Pro}(X_2 \mid \delta_A^{PL}(x)) = 0.25$. In other words, there is no conflict between such two different lower approximations.

The above example tells us that our pseudolabel neighborhood approach can not only provide us with better classification performance, but also remove the conflict case between the results of lower approximations.

Similar to what have been addressed for NDTRS in [16], we can also obtain the following results for PLNDTRS.

Proposition 11. *Given a pseudolabel decision system DS^{PL} and two values α_1, α_2, if $\alpha_1 \leq \alpha_2$, then $\forall A \subseteq AT$ and $\forall X \subseteq U$; we have* $\text{PLPOS}_A^{(\alpha_2,\beta)}(X) \subseteq \text{PLPOS}_A^{(\alpha_1,\beta)}(X)$.

Proof. $\forall y \in \text{PLPOS}_A^{(\alpha_2,\beta)}(X)$, by Definition 9 and Equation (19), we have $\text{Pro}(X \mid \delta_A^{PL}(y)) \geq \alpha_2$. Since $\alpha_1 \leq \alpha_2$, we then have $\text{Pro}(X \mid \delta_A^{PL}(y)) \geq \alpha_1$; it follows that $y \in \text{PLPOS}_A^{(\alpha_1,\beta)}(X)$; that is, $\text{PLPOS}_A^{(\alpha_2,\beta)}(X) \subseteq \text{PLPOS}_A^{(\alpha_1,\beta)}(X)$ is proved. $\square$

Proposition 12. *Given a pseudolabel decision system DS^{PL} and two values β_1, β_2, if $\beta_1 \leq \beta_2$, then $\forall A \subseteq AT$ and $\forall X \subseteq U$; we have* $\text{PLBND}_A^{(\alpha,\beta_2)}(X) \subseteq \text{PLBND}_A^{(\alpha,\beta_1)}(X)$.

Proof. For $\forall y \in \text{PLBND}_A^{(\alpha,\beta_2)}(X)$, by Definition 9, we have $\text{Pro}(X \mid \delta_A^{PL}(y)) > \beta_2$. Since $\beta_1 \leq \beta_2$, we then have $\text{Pro}(X \mid \delta_A^{PL}(y)) > \beta_1$; it follows that $y \in \text{PLBND}_A^{(\alpha,\beta_1)}(X)$; that is, $\text{PLBND}_A^{(\alpha,\beta_2)}(X) \subseteq \text{PLBND}_A^{(\alpha,\beta_1)}(X)$ is proved. $\square$

The above two propositions suggest that we can also adjust the size of positive region or the boundary region of X by modifying the values of α and β in PLNDTRS.

4. Attribute Reduction in PLNDTRS

4.1. Aim of Attribute Reduction. Attribute reduction [43] plays a crucial role in the development of rough set. Different from previous feature selections in the field of Machine Learning, attribute reduction has clear explanations with respect to different constraints. For example, if the positive region is expected to be preserved, then attribute reduction may be defined as a minimal subset of the condition attributes which preserves the positive region.

Up to now, through considering various constraints, a lot of the researchers have proposed so many different definitions of attribute reductions [43–48]. It should be noticed that, for DTRS related attribute reductions, decision costs provide us with an interesting topic [11, 49, 50]. In other words, it is frequently expected that, through using attribute reduction, the obtained subset of the condition attributes may provide lower or minimal value of decision cost. Therefore, in the following of this section, the attribute reduction will be further explored for achieving lower decision cost based on PLNDTRS.

Firstly, given a pseudolabel decision system DS^{PL}, let us review how to calculate the decision costs. $\forall A \subseteq AT$ and suppose that $U/\text{IND}(d) = \{X_1, X_2, \ldots, X_q\}$; the total decision cost is

$$\mathbf{COST}_A^{PL} = \text{COST}_A^{\text{PLPOS}} + \text{COST}_A^{\text{PLBND}} \\ + \text{COST}_A^{\text{PLNEG}}; \tag{20}$$

where

$$\text{COST}_A^{\text{PLPOS}} = \sum_{j=1}^{q} \sum_{x \in \text{PLPOS}_A^{(\alpha,\beta)}(X_j)} \left(\text{Pro}\left(X_j \mid \delta_A^{PL}(x) \right) \right. \\ \left. \cdot \lambda_{PP} + \left(1 - \text{Pro}\left(X_j \mid \delta_A^{PL}(x) \right) \right) \cdot \lambda_{PN} \right); \tag{21}$$

TABLE 2: A small example of pseudolabel decision system.

	a_1	a_2	a_3	a_4	a_5	d	d^{PL}
x_1	0.5402	0.8300	0.4755	0.0212	0.4771	1	#
x_2	0.9538	0.4890	0.5832	0.8799	0.6534	1	#
x_3	0.2089	0.7607	0.2605	0.7981	0.9665	2	#
x_4	0.1163	0.9151	0.0848	0.3241	0.3130	1	#
x_5	0.6462	0.9009	0.2981	0.6690	0.0764	3	#
x_6	0.1084	0.2142	0.9171	0.2962	0.7914	3	*
x_7	0.9834	0.5470	0.4705	0.9299	0.3653	2	#
x_8	0.2483	0.7847	0.2694	0.2819	0.5850	1	#
x_9	0.6063	0.1944	0.7629	0.1688	0.1833	2	+
x_{10}	0.8166	0.7468	0.7721	0.7451	0.0769	2	*

$$\text{COST}_A^{\text{PLBND}} = \sum_{j=1}^{q} \sum_{x \in \text{PLBND}_A^{(\alpha,\beta)}(X_j)} \left(\text{Pro}\left(X_j \mid \delta_A^{PL}(x)\right) \right. \tag{22}$$

$$\left. \cdot \lambda_{BP} + \left(1 - \text{Pro}\left(X_j \mid \delta_A^{PL}(x)\right)\right) \cdot \lambda_{BN} \right);$$

$$\text{COST}_A^{\text{PLNEG}} = \sum_{j=1}^{q} \sum_{x \in \text{PLNEG}_A^{(\alpha,\beta)}(X_j)} \left(\text{Pro}\left(X_j \mid \delta_A^{PL}(x)\right) \right. \tag{23}$$

$$\left. \cdot \lambda_{NP} + \left(1 - \text{Pro}\left(X_j \mid \delta_A^{PL}(x)\right)\right) \cdot \lambda_{NN} \right).$$

$\textbf{COST}_A^{PL}$ indicates the total decision cost, $\text{COST}_A^{\text{PLPOS}}$ indicates the decision cost of positive region, $\text{COST}_A^{\text{PLBND}}$ indicates the decision cost of boundary region, and $\text{COST}_A^{\text{PLNEG}}$ is the decision cost of negative region.

Similar to other DTRS, the decision costs of our PLNDTRS are not always monotonic if the used condition attributes vary. For example, if the number of used condition attributes has been reduced, then the derived decision costs may increase or decrease. Therefore, our definition attribute reduction will be presented, which aims to achieve the lower decision cost through selecting a subset of the condition attributes.

Definition 13. Given a pseudolabel decision system DS^{PL}, $\forall A \subseteq AT$, A is referred to as a reduct in DS^{PL} if and only if

(1) $\textbf{COST}_A^{PL} \leq \textbf{COST}_{AT}^{PL}$ and $A \neq \emptyset$;

(2) $\forall B \subset A$, $\textbf{COST}_B^{PL} > \textbf{COST}_A^{PL}$.

Example 14. Table 2 is a small example of pseudolabel decision system which contains 10 samples. All of these samples are described by five condition attributes such that $AT = \{a_1, a_2, a_3, a_4, a_5\}$. The pseudolabels of samples are obtained by clustering technique.

Assuming that $\delta = 0.2$ and the parameters in loss function are $\lambda_{PP} = \lambda_{NN} = 0$, $\lambda_{PN} = 1$, $\lambda_{NP} = 0.95$, $\lambda_{BP} = 0.16$, $\lambda_{BN} = 0.39$, then we can calculate two thresholds such that $\alpha = 0.7922$ and $\beta = 0.3305$. The decision cost based on the whole condition attribute set AT, that is, $\textbf{COST}_{AT}^{PL}$, is 1.2983.

By computation, the decision cost based on $\{a_3, a_5\}$ is 0.8633, which is lower than $\textbf{COST}_{AT}^{PL}$. Furthermore, if we

remove the attribute a_5, then the decision cost based on $\{a_3\}$ is 2.0017; if we remove the attribute a_3, then the decision cost based on $\{a_5\}$ is 1.4983. Obviously, both the decision cost based on $\{a_3\}$ and the decision cost based on $\{a_5\}$ are higher than the decision cost based on $\{a_3, a_5\}$; that is, neither a_3 nor a_5 can be removed for generating lower decision cost. Therefore, the condition attributes set $\{a_3, a_5\}$ is a reduct in Table 2.

4.2. Algorithm to Compute Reduct. Presently, in the field of rough set, it is well known that obtaining all reducts is an NP-hard problem [51, 52]; it follows that many heuristic algorithms for finding reducts have been studied [45, 53]. Generally speaking, a heuristic algorithm contains two core aspects: heuristic information and searching strategy.

In the heuristic algorithm, the fitness function can be used to characterize the heuristic information. For example, in Definition 13, the decision cost $\textbf{COST}_A^{PL}$ should be used to construct fitness function because the aim of Definition 13 is to find a subset of condition attributes which derives lower decision cost.

With regard to the searching strategies in the heuristic algorithm, two approaches have been considered. One is directional searching and the other is nondirectional searching. The directional searching strategy can be further categorized into deletion method, addition method, and addition-deletion method [54]. Nondirectional searching strategy is usually applied to evolutionary algorithms, swarm algorithms, and other population-based metaheuristic algorithms for optimization problems.

In the context of this paper, we will use the addition method to compute the reduct; that is, the searching strategy is directional searching. To achieve that, the fitness function for measuring the significance of the condition attribute should be presented. The detailed form is shown in the following definition.

Definition 15. Given a pseudolabel decision system DS^{PL}, $\forall A \subseteq AT$, and $\forall a \in AT - A$, the fitness value of attribute a is

$$\text{Sig}(a, A, d) = \textbf{COST}_{A \cup \{a\}}^{PL} - \textbf{COST}_A^{PL}. \tag{24}$$

Inputs: A decision system DS.
Outputs: A reduct A.
Step 1: $\forall x_i \in U$, obtain $d_{AT}^{PL}(x_i)$ by k-means clustering.
    $//k$ is same to the number of decision classes in DS.
Step 2: Compute $\mathbf{COST}_{AT}^{PL}$;
Step 3: $A = \emptyset$ and let $\mathbf{COST}_A^{PL}$ be a value which is greater than $\mathbf{COST}_{AT}^{PL}$;
Step 4: While $\mathbf{COST}_A^{PL} > \mathbf{COST}_{AT}^{PL}$;
    (1) $\forall AT - A$, $\forall x_i \in U$, compute $d_{A \cup \{a\}}^{PL}(x_i)$;
    (2) $\forall AT - A$, compute $\mathrm{Sig}(a, A, d)$;
    (3) Select the attribute a such that $a = \arg\min\{\mathrm{Sig}(a, A, d) : \forall a \in AT - A\}$;
    (4) $A = A \cup \{a\}$;
    (5) Compute $\mathbf{COST}_A^{PL}$;
    End While.
Step 5: Output A.

ALGORITHM 1: Compute reduct based on PLNDTRS.

TABLE 3: A description of data sets.

ID	Data sets	Samples	Attributes	Decision classes
1	Abalone	4177	7	29
2	Breast Cancer Wisconsin (Diagnostic)	569	30	2
3	Caffeine Consumption	1885	12	7
4	Cardiotocography	1216	21	10
5	Dermatology	366	36	6
6	E. coli	336	7	8
7	Glass Identification	214	9	6
8	LSVT Voice Rehabilitation	126	257	2
9	Page Blocks Classification	5473	10	5
10	Ring	7400	20	2
11	Statlog (Image Segmentation)	2310	18	7
12	Urban Land Cover	675	147	9
13	Waveform Database Generator	5000	21	3
14	Wine	178	13	3
15	Yeast	1484	8	10

Obviously, the above fitness function displays the significance or importance of an attribute. If $\mathrm{Sig}(a, A, d) \geq 0$, then the attribute a cannot be added into A because the decision cost will be increased; if $\mathrm{Sig}(a, A, d) < 0$, then the attribute a can be added into A possibly because the decision cost will be decreased.

In the procedure of adding attributes, if $\mathbf{COST}_A^{PL} \leq \mathbf{COST}_{AT}^{PL}$, then we will stop such procedure and output the final reduct A. The detailed process is shown in Algorithm 1 [16, 49].

In Algorithm 1, following the basic structure of k-means clustering, in Step 1, the time complexity of computing pseudolabels of samples is $O(k \cdot T \cdot |U|)$, in which k is the number of clusters and T is the number of iterations. The time complexity of Step 2 is $O(|U|^2 \cdot |AT|)$ because it is required to compare any two samples in U over condition attributes AT for generating neighborhoods. Step 4 is the iteration process of adding attributes into the reduct. The time complexity of Step 4 is $O(|U|^2 \cdot |AT|^2)$ since, in the worst case, such

looping should be executed $|AT|$ times. To sum up, we can conclude that the whole time complexity of Algorithm 1 is $O(k \cdot T \cdot |U| + |U|^2 \cdot |AT|^2)$.

5. Experimental Analyses

In this section, to evaluate the performance of PLNDTRS and corresponding algorithm to compute reduct, 15 UCI data sets have been selected to conduct the experiments. The details of these data are shown in Table 3. All the experiments have been carried out on a personal computer with Windows 10, Intel Core i5 3230M CPU (2.60GHz), and 4.00 GB memory. The programming platform is MATLAB R2018a.

In the experiments, k-means clustering is employed to produce pseudolabels, and the value of k is the number of decision classes. For all the experiments in this section, 10 different values of δ have been used; they are $\delta = 0.03, 0.06, \ldots, 0.3$. Moreover, not only the 10-fold cross-validation is employed in our experiments, but also 10

TABLE 4: Comparisons of different decision costs in terms of raw attributes (smaller values are underlined).

	Total costs		Costs of positive region		Costs of boundary region		Costs of negative region	
ID	NDTRS	PLNDTRS	NDTRS	PLNDTRS	NDTRS	PLNDTRS	NDTRS	PLNDTRS
1	515.7637	505.4116	3.1993	4.0853	466.0496	448.8289	46.5148	52.4974
2	50.8504	26.4840	31.3946	23.0951	19.4558	3.3889	0.0000	0.0000
3	399.2153	388.7867	100.7310	93.6039	297.0621	293.5297	1.4220	1.6532
4	424.4182	397.1313	31.8888	33.7502	357.4530	336.0896	35.0765	27.2915
5	22.3292	6.4481	5.3326	4.2677	16.8490	2.1804	0.1475	0.0000
6	49.7287	36.9805	13.7807	11.7077	35.2870	24.8976	0.6611	0.3752
7	48.5359	45.3767	2.1099	1.7737	45.7020	43.1462	0.7239	0.4568
8	28.0053	27.6984	2.3890	2.3280	25.6164	25.3704	0.0000	0.0000
9	247.8693	221.4285	176.073	149.7464	70.7042	71.6207	0.9431	0.0613
10	640.0481	634.8294	364.5238	360.2211	275.5243	274.6083	0.0000	0.0000
11	272.5914	249.7893	43.8713	41.2862	214.8033	206.0070	13.9169	2.4961
12	131.3945	97.1666	20.5268	26.6987	107.7290	69.9520	3.1387	0.5159
13	967.5386	944.4428	209.2039	194.6551	758.3347	749.7877	0.0000	0.0000
14	8.7816	3.9627	4.8234	2.3965	3.9583	1.5662	0.0000	0.0000
15	381.1040	361.9059	11.7836	21.0489	319.6782	315.4133	49.6422	25.4437

different groups of loss functions are generated randomly for each cross-validation. Therefore, for each radius δ, each data set will be tested 100 times.

5.1. Experiments on Raw Attributes. In what follows, the decision costs obtained by NDTRS and PLNDTRS in terms of raw attributes will be compared. The forms of these decision costs have been shown in Equations (20)-(23). The detailed results are shown in Table 4.

Through a careful investigation of Table 4, it is not difficult to draw the following conclusion.

For all the data sets, the total decision costs generated by PLNDTRS are lower than those generated by NDTRS. This is mainly because (1) most of the decision costs of PLNDTRS based positive region are lower than those of NDTRS based positive region; (2) most of the decision costs of PLNDTRS based boundary region are lower than those of NDTRS based boundary region. From this point of view, by comparing NDTRS, our pseudolabel strategy does work for reducing the decision costs.

5.2. Experiments on Attribute Reduction. In this subsection, to show the efficiency of the proposed attribute reduction shown in Definition 13, the decision costs derived by the obtained reducts will be compared.

Figure 4 shows us the detailed change trend lines of decision costs. It should be noticed that the decision costs shown in Figure 4 are total decision costs instead of the decision costs of three different regions. In each subfigure of Figure 4, the x-coordinate pertains to value of δ, whereas the y-coordinate concerns the values of costs.

The detailed legends used in Figure 4 are (1) Raw-Attributes: NDTRS based on the original attributes; (2) NDTRS-Reduct: NDTRS based on the attribute reduction proposed in [16] (similar to our Definition 13, the aim of this attribute reduction is also to find a minimal subset of the condition attributes which can decrease the total decision

cost); (3) PLRaw-Attributes: PLNDTRS based on the original attributes; (4) PLNDTRS-Reduct: PLNDTRS based on the attribute reduction proposed in Definition 13.

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

(1) By comparing with Raw-Attributes, PLRaw-Attributes may offer the lower decision costs. This observation is consistent with what have been addressed in Section 5.1; that is, our pseudolabel approach does decrease the total decision costs.

(2) By comparing with NDTRS-Reduct, PLNDTRS-Reduct also provides us with the lower decision costs. In other words, through attribute reductions, our pseudolabel approach can find subsets of attributes which offer smaller decision costs though the aims of attribute reductions based on NDTRS and PLNDTRS are the same.

(3) In most data sets, both NDTRS-Reduct and PLNDTRS-Reduct do reduce the decision costs. That is to say, both the attribute reductions proposed by Li et al. in [16] and us are effective.

(4) With the increasing of value of δ, in most data sets, four different types of decision costs will increase. This is mainly because if the value of δ is greater, then the size of the neighborhood will be larger; it follows that the positive region shows shrinking and the negative region will be expanded. Correspondingly, the decision costs with respect to different regions will be changed.

Furthermore, the decision costs related to three different regions are shown in Table 5.

With a careful investigation of Table 5, we notice that the reducing of total decision costs shown in Figure 4 mainly comes from the reducing of costs of positive region and costs
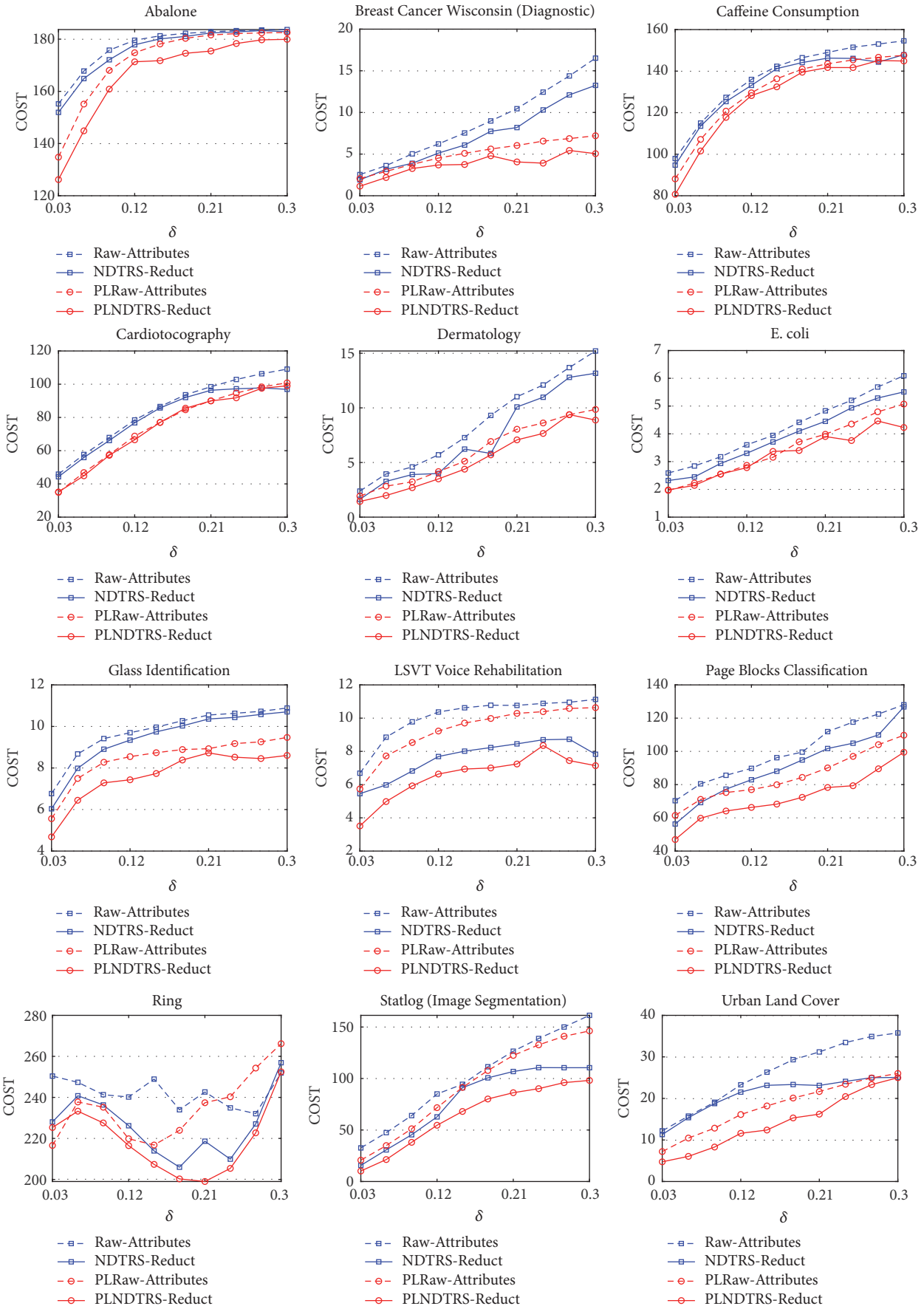
Abalone

Breast Cancer Wisconsin (Diagnostic)

Caffeine Consumption

- -□- Raw-Attributes
- —□— NDTRS-Reduct
- -○- PLRaw-Attributes
- —○— PLNDTRS-Reduct

Cardiotocography

Dermatology

E. coli

- -□- Raw-Attributes
- —□— NDTRS-Reduct
- -○- PLRaw-Attributes
- —○— PLNDTRS-Reduct

Glass Identification

LSVT Voice Rehabilitation

Page Blocks Classification

- -□- Raw-Attributes
- —□— NDTRS-Reduct
- -○- PLRaw-Attributes
- —○— PLNDTRS-Reduct

Ring

Statlog (Image Segmentation)

Urban Land Cover

- -□- Raw-Attributes
- —□— NDTRS-Reduct
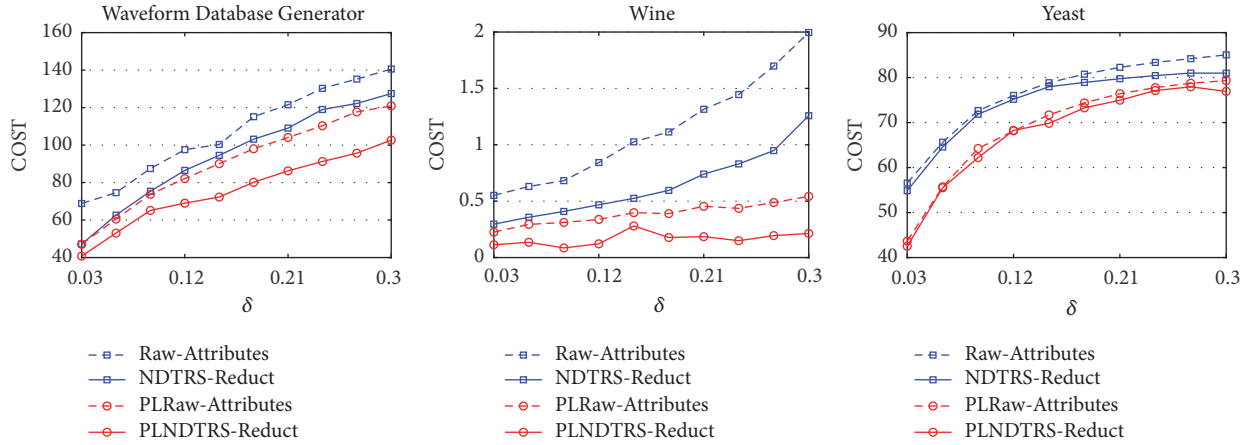- -○- PLRaw-Attributes
- —○— PLNDTRS-Reduct

FIGURE 4: Continued.

Figure 4: A comparison among different decision costs.

Table 5: Decision costs related to three different regions (smaller values are underlined).

ID	Costs of positive region		Costs of boundary region		Costs of negative region	
	NDTRS-Reduct	PLNDTRS-Reduct	NDTRS-Reduct	PLNDTRS-Reduct	NDTRS-Reduct	PLNDTRS-Reduct
1	3.3083	2.5927	144.6449	137.0848	28.6878	28.2077
2	3.5871	2.7336	3.5886	0.9952	0.0000	0.0000
3	27.4517	22.2394	47.4321	45.5593	0.4478	0.5550
4	4.4652	5.6460	53.4615	57.5998	22.9725	11.2165
5	1.3671	1.4976	5.5215	3.5962	0.2970	0.1747
6	0.8422	0.5908	2.9842	2.6384	0.0727	0.0283
7	1.4726	1.0331	8.6575	6.3638	0.2810	0.2297
8	0.5302	0.5050	0.9215	0.8607	0.0000	0.0000
9	48.6126	40.8185	32.7738	30.1912	0.0378	0.1700
10	141.8567	104.5482	79.8204	132.0821	0.0000	0.0000
11	9.9494	9.8133	26.6506	33.9288	44.3499	33.8493
12	1.1821	3.7938	4.7748	11.7701	14.6167	1.5715
13	19.0932	17.6336	73.3651	77.0279	0.0276	0.0451
14	0.2434	0.0585	0.3966	0.1073	0.0029	0.0000
15	2.1211	3.3493	37.0947	50.0695	35.3422	14.4422

of boundary region. This result is consistent with what have been addressed in Section 5.1.

Furthermore, to further demonstrate the effectiveness of our attribute reduction based on PLNDTRS, we will compare it with other popular attribute reduction methods in DTRS; they are (1) POS-Reduct (positive region extension based attribute reduction [8, 14, 49]) and NON-NEG-Reduct (non-negative region extension based attribute reduction [49]). The details are shown in Figure 5.

By Figure 5, it is not difficult to observe the following.

(1) Generally speaking, if the value of δ increases, then the decision costs obtained by four different reducts increase. In other words, the performances of different reducts are closely related to the scales of radii.

(2) In most cases, the decision costs derived from "PLNDTRS-Reduct" are lower than those derived

from "NDTRS-Reduct," "POS-Reduct," and "NON-NEG-Reduct." Take "Yeast" data set as an example; if δ = 0.12, then the decision cost related to "PLNDTRS-Reduct" is 105 while the decision cost related to "NDTRS-Reduct" is 120; the decision cost related to "POS-Reduct" is 121 and the decision cost related to "NON-NEG-Reduct" is 138. From this point of view, we can say that our pseudolabel strategy is superior to several previous research results about attribute reduction in DTRS.

Finally, we will also compare the lengths of different reducts and the time consumptions for computing these reducts. The details are shown in Tables 6 and 7.

Following Tables 6 and 7, it is not difficult to observe that more time is required to compute PLNDTRS-reduct. The main reason includes two aspects: (1) for each iteration in computing reduct, the pseudolabels of samples should be regenerated; this has been pointed out in (1) of Step 4
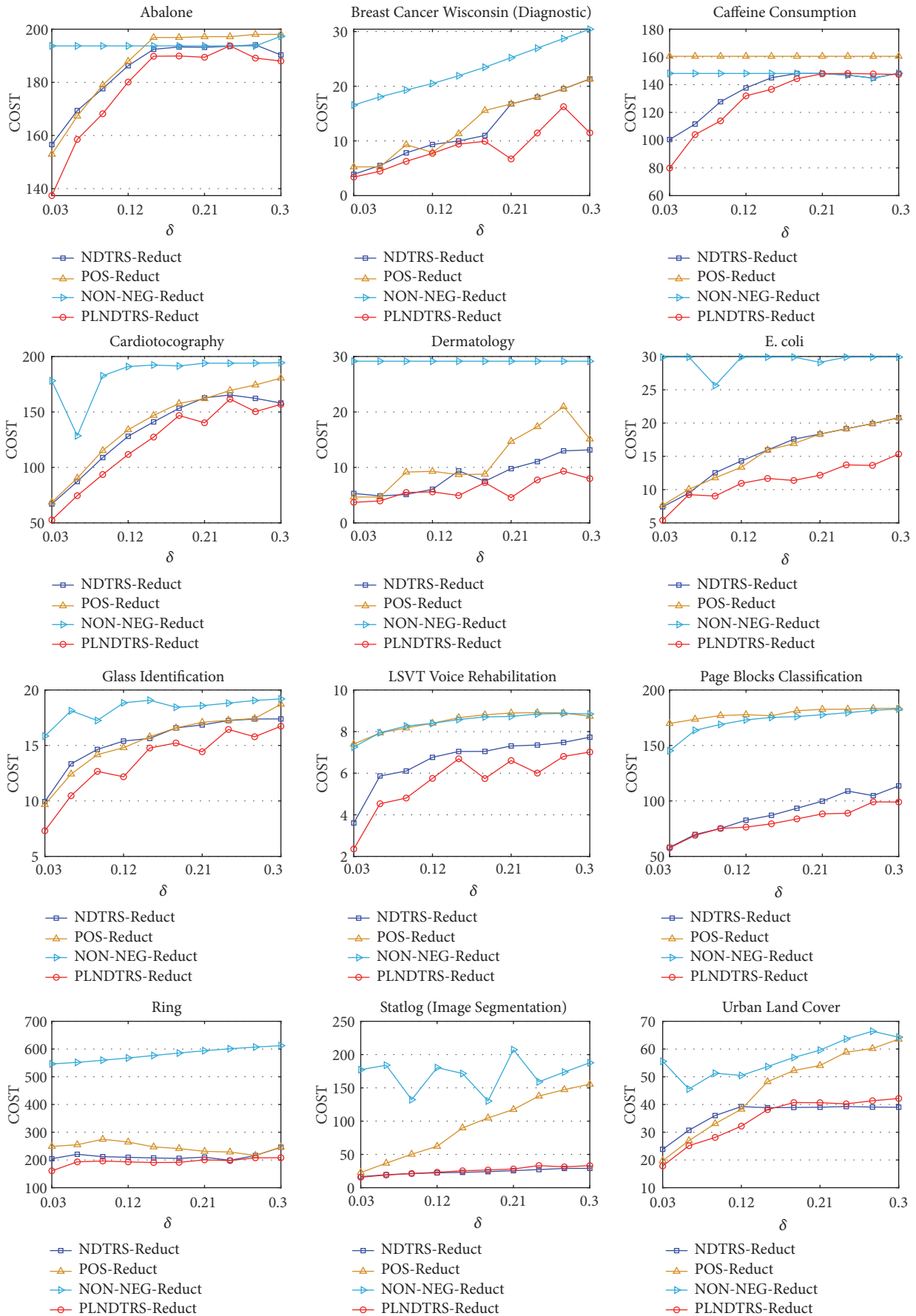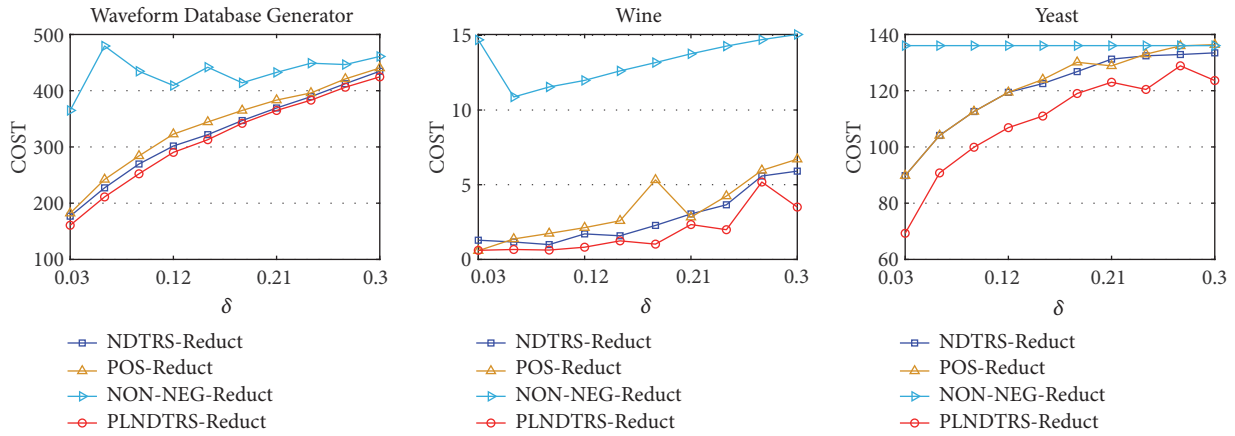
FIGURE 5: Continued.

FIGURE 5: A comparison among decision costs derived by different reducts.

TABLE 6: The average lengths of different reducts (shorter lengths are underlined).

ID	NON-NEG-Reduct	POS-Reduct	NDTRS-Reduct	PLNDTRS-Reduct
1	1.00	2.56	2.76	3.50
2	1.00	2.10	1.80	3.20
3	1.00	1.37	2.40	2.80
4	1.70	7.00	9.40	10.23
5	1.00	6.00	5.77	8.13
6	1.00	3.77	3.30	5.23
7	1.00	3.90	4.20	3.33
8	1.00	1.00	1.03	1.03
9	1.00	1.00	4.06	6.16
10	1.00	16.50	17.76	17.80
11	2.50	4.13	7.00	6.76
12	1.97	4.37	12.47	37.80
13	2.70	6.86	8.73	8.83
14	1.00	4.53	4.67	5.03
15	2.70	4.30	3.70	5.50

TABLE 7: The time consumptions (seconds) for computing different reducts (lower consumptions are underlined).

ID	NON-NEG-Reduct	POS-Reduct	NDTRS-Reduct	PLNDTRS-Reduct
1	27.7031	29.4720	97.7107	182.7403
2	1.2874	1.3160	1.8312	3.2800
3	2.6965	2.7879	8.5677	11.7441
4	10.4468	18.8002	61.4017	81.0210
5	0.1894	0.5577	0.5562	6.0147
6	0.0452	0.0657	0.1860	0.5958
7	0.0234	0.0366	0.1176	0.3378
8	1.1789	1.0911	1.2891	2.3478
9	48.4413	24.2008	241.5598	344.9764
10	243.8521	1146.7207	3521.3002	3619.4430
11	15.8729	11.9315	611.5846	915.7841
12	5.2911	6.0499	45.5096	271.2546
13	188.2825	217.8035	791.8680	830.0859
14	0.0215	0.0449	0.1320	0.4860
15	2.0325	1.4747	3.7970	7.3192

TABLE 8: p-values of Wilcoxon signed rank test for comparing reducts (higher p-values are italic).

	Total costs	Costs of positive region	Costs of boundary region	Costs of negative region
ID	NDTRS-Reduct & PLNDTRS-Reduct	NDTRS-Reduct & PLNDTRS-Reduct	NDTRS-Reduct & PLNDTRS-Reduct	NDTRS-Reduct & PLNDTRS-Reduct
1	0.0020	*0.0977*	0.0273	*0.9219*
2	0.0020	0.0020	0.0020	*1.0000*
3	0.0020	0.0098	*0.6953*	*0.3750*
4	0.0059	*0.2754*	*0.5566*	0.0098
5	0.0020	*0.9219*	0.0195	*0.1563*
6	0.0020	0.0098	0.0020	*0.1309*
7	0.0020	*0.0645*	0.0020	*0.2754*
8	0.0020	0.0039	0.0400	*1.0000*
9	0.0098	0.0020	0.0050	0.0095
10	0.0488	0.0039	0.0059	*1.0000*
11	0.0316	*0.0695*	0.0050	0.0039
12	0.0137	0.0371	0.0195	0.0020
13	0.0195	0.0020	0.0137	*1.0000*
14	0.0020	0.0078	0.0020	*1.0000*
15	0.0020	0.0488	*0.1993*	0.0020

in Algorithm 1; (2) based on Table 6, more attributes are required to construct reducts; such fact indicates that more iterations should be executed.

Moreover, it must be emphasized that though the time consumptions of calculating NON-NEG-Reduct and POS-Reduct are lower, such two types of reducts may not be good enough for providing smaller values of decision costs. This fact can be observed in Figure 5 clearly. From this point of view, our pseudolabel strategy based attribute reduction is superior to previous researches though our approach requires more time to obtain reduct.

5.3. Statistical Comparisons of Reducts. In this section, we will make the statistical comparisons between PLNDTRS-Reduct and NDTRS-Reduct. The Wilcoxon signed rank test [55] will be selected for comparing such two reducts. The purpose of this computation is trying to reject the null hypothesis that the two reducts perform equally well.

For each data set, we have used 10 different radii to obtain reducts; it follows that 10 decision costs will be generated by each algorithm. Take the data "Caffeine Consumption" for instance, the 10 total decision costs derived from NDTRS-Reduct are "54.9455, 60.6065, 70.4899, 75.1574, 77.0271, 81.4503, 83.2687, 82.8129, 82.6185, 84.9391" while the 10 total decision costs derived from PLNDTRS-Reduct are "47.6352, 57.7943, 63.3378, 69.9135, 73.6124, 76.0755, 78.9889, 81.4714, 81.2164, 83.4922;" the corresponding p-value of Wilcoxon signed rank test is then 0.002. 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 8.

Following the results of Table 8, if the significance level is given by 0.05, we therefore reject the null hypothesis. In Table 8, we notice that most of the p-values are lower than 0.05. In other words, from the viewpoint of costs, reducts

based on NDTRS and those based on our PLNDTRS do not perform equally well though the aim of such two reducts is the same.

Remark 16. Most of the p-values in "costs of negative region" are equal to 1. This is mainly because, based on NDTRS-Reduct and PLNDTRS-Reduct, the decision costs of negative region are 0 in most cases.

6. Conclusions and Further Perspectives

By considering the label information of samples, a framework of pseudolabel strategy has been introduced into the model of decision-theoretic rough set. Different from the traditional constructions of decision-theoretic rough set, our approach is achieved by not only the distance based neighborhood, but also the pseudolabels of samples. The experimental results have demonstrated that our pseudolabel approach can reduce the decision costs which are closely related to decision-theoretic rough set. Moreover, the attribute reduction based on our pseudolabel strategy can also provide attributes with better performances if decision costs are taken into consideration.

The following topics are challenges for further research.

(1) Only the k-means clustering approach is used to generate pseudolabels of samples; the label propagation or supervised approach will be further explored.

(2) Through using pseudolabel strategy, how to design quick process to compute reduct for large scale data [56–58] is another interesting topic to be addressed.

Data Availability

The [UCI] data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

[1] Y. Qian, X. Liang, Q. Wang et al., "Local rough set: a solution to rough data analysis in big data," *International Journal of Approximate Reasoning*, vol. 97, pp. 38–63, 2018.

[2] J. Ma, H. Zhang, and Y. Qian, "Three-way decisions with reflexive probabilistic rough fuzzy sets," *Granular Computing*, vol. 1, pp. 1–13, 2018.

[3] W. Ma, L. Yue, B. Sun, and H. Zhao, "Parameter selection and uncertainty measurement for variable precision probabilistic rough set," *Computing and Informatics*, vol. 37, no. 3, pp. 614–634, 2018.

[4] Y. Y. Yao and S. K. M. Wong, "A decision theoretic framework for approximating concepts," *International Journal of Man-Machine Studies*, vol. 37, no. 6, pp. 793–809, 1992.

[5] Y. Yao, "Three-way decisions with probabilistic rough sets," *Information Sciences*, vol. 180, no. 3, pp. 341–353, 2010.

[6] Y. Yao, "An outline of a theory of three-way decisions," in *Rough Sets and Current Trends in Computing*, vol. 7413 of *Lecture Notes in Computer Science*, pp. 1–17, Springer, Berlin, Heidelberg, Germany, 2012.

[7] H. Chen, T. Li, C. Luo, S.-J. Horng, and G. Wang, "A decision-theoretic rough set approach for dynamic data mining," *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 6, pp. 1958–1970, 2015.

[8] H. Ju, H. Li, X. Yang, X. Zhou, and B. Huang, "Cost-sensitive rough set: a multi-granulation approach," *Knowledge-Based Systems*, vol. 123, pp. 137–153, 2017.

[9] P. Mandal and A. S. Ranadive, "Decision-theoretic rough sets under pythagorean fuzzy information," *International Journal of Intelligent Systems*, vol. 33, no. 4, pp. 818–835, 2018.

[10] H. Dou, X. Yang, X. Song, H. Yu, W.-Z. Wu, and J. Yang, "Decision-theoretic rough set: a multicost strategy," *Knowledge-Based Systems*, vol. 91, pp. 71–83, 2016.

[11] X. Y. Jia, Z. M. Tang, W. H. Liao, and L. Shang, "On an optimization representation of decision-theoretic rough set model," *International Journal of Approximate Reasoning*, vol. 55, no. 1, part 2, pp. 156–166, 2014.

[12] L. A. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets and Systems*, vol. 90, no. 2, pp. 111–127, 1997.

[13] D. Liu and D. Liang, "Three-way decisions in ordered decision system," *Knowledge-Based Systems*, vol. 137, pp. 182–195, 2017.

[14] H. Ju, X. Yang, H. Yu, T. Li, D.-J. Yu, and J. Yang, "Cost-sensitive rough set approach," *Information Sciences*, vol. 355-356, pp. 282–298, 2016.

[15] J. Song, E. C. C. Tsang, D. Chen, and X. Yang, "Minimal decision cost reduct in fuzzy decision-theoretic rough set model," *Knowledge-Based Systems*, vol. 126, pp. 104–112, 2017.

[16] W. Li, Z. Huang, X. Jia, and X. Cai, "Neighborhood based decision-theoretic rough set models," *International Journal of Approximate Reasoning*, vol. 69, pp. 1–17, 2016.

[17] D. Liu, T. R. Li, and H. X. Li, "A multiple-category classification approach with decision-theoretic rough sets," *Fundamenta Informaticae*, vol. 115, no. 2-3, pp. 173–188, 2012.

[18] B. Zhou, "Multi-class decision-theoretic rough sets," *International Journal of Approximate Reasoning*, vol. 55, no. 1, part 2, pp. 211–224, 2014.

[19] Y. H. Qian, H. Zhang, Y. L. Sang, and J. Y. Liang, "Multigranulation decision-theoretic rough sets," *International Journal of Approximate Reasoning*, vol. 55, no. 1, part 2, pp. 225–237, 2014.

[20] X. Jia, W. Li, and L. Shang, "A multiphase cost-sensitive learning method based on the multiclass three-way decision-theoretic rough set model," *Information Sciences*, vol. 485, pp. 248–262, 2019.

[21] W. Li, X. Jia, L. Wang, and B. Zhou, "Multi-objective attribute reduction in three-way decision-theoretic rough set model," *International Journal of Approximate Reasoning*, vol. 105, pp. 327–341, 2019.

[22] D. C. Liang and D. Liu, "Systematic studies on three-way decisions with interval-valued decision-theoretic rough sets," *Information Sciences*, vol. 276, pp. 186–203, 2014.

[23] X. Yang, T. Li, H. Fujita, and D. Liu, "A sequential three-way approach to multi-class decision," *International Journal of Approximate Reasoning*, vol. 104, pp. 108–125, 2019.

[24] J. Zhang, T. Li, D. Ruan, and D. Liu, "Neighborhood rough sets for dynamic data mining," *International Journal of Intelligent Systems*, vol. 27, no. 4, pp. 317–342, 2012.

[25] L. Sun, X. Zhang, J. Xu, and S. Zhang, "An attribute reduction method using neighborhood entropy measures in neighborhood rough sets," *Entropy*, vol. 21, no. 2, pp. 155–177, 2019.

[26] C. Wang, Y. Shi, X. Fan, and M. Shao, "Attribute reduction based on k-nearest neighborhood rough sets," *International Journal of Approximate Reasoning*, vol. 106, pp. 18–31, 2019.

[27] Q. Hu, D. Yu, and Z. Xie, "Neighborhood classifiers," *Expert Systems with Applications*, vol. 34, no. 2, pp. 866–876, 2008.

[28] S. Xu, X. Yang, E. C. C. Tsang, and E. A. Mantey, "Neighborhood collaborative classifiers," in *Proceedings of the 2016 International Conference on Machine Learning and Cybernetics, ICMLC 2016*, pp. 470–476, Republic of Korea, July 2016.

[29] T. Denœux, S. Sriboonchitta, and O. Kanjanatarakul, "Evidential clustering of large dissimilarity data," *Knowledge-Based Systems*, vol. 106, pp. 179–195, 2016.

[30] J. Tang and 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, KDD 2012*, pp. 904–912, China, August 2012.

[31] J. Tang, X. Hu, H. Gao, and H. Liu, "Discriminant analysis for unsupervised feature selection," in *Proceedings of the 14th SIAM International Conference on Data Mining, SDM 2014*, pp. 938–946, USA, April 2014.

[32] K. Wu and K.-H. Yap, "Fuzzy SVM for content-based image retrieval," *IEEE Computational Intelligence Magazine*, vol. 1, no. 2, pp. 10–16, 2006.

[33] W. Zeng, X. Chen, and H. Cheng, "Pseudo labels for imbalanced multi-label learning," in *Proceedings of the 2014 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2014*, pp. 25–31, China, November 2014.

[34] P. Zhu, W. Zhu, Q. Hu, C. Zhang, and W. Zuo, "Subspace clustering guided unsupervised feature selection," *Pattern Recognition*, vol. 66, pp. 364–374, 2017.

[35] W. Pedrycz, "Granular computing: an introduction," in *Proceedings of the Ifsa World Congress and Nafips International Conference*, pp. 1349–1354, 2002.

[36] J. Liang and Y. Qian, "Information granules and entropy theory in information systems," *Science China Information Sciences*, vol. 51, no. 10, pp. 1427–1444, 2008.

[37] Y. Yao and X. Deng, "Sequential three-way decisions with probabilistic rough sets," *Information Sciences*, vol. 180, no. 3, pp. 341–353, 2010.

[38] Z. Pawlak, "Granularity of knowledge, indiscernibility and rough sets," in *Proceedings of the 1998 IEEE International Conference on Fuzzy Systems, FUZZY 1998*, pp. 106–110, USA, May 1998.

[39] S. Liang, K. Liu, X. Chen, P. Wang, and X. Yang, "Multi-granularity attribute reduction," in *International Joint Conference on Rough Sets*, pp. 61–72, 2018.

[40] J. Li, C. Huang, J. Qi, Y. Qian, and W. Liu, "Three-way cognitive concept learning via multi-granularity," *Information Sciences*, vol. 378, pp. 244–263, 2017.

[41] P. Zhu and Q. Hu, "Adaptive neighborhood granularity selection and combination based on margin distribution optimization," *Information Sciences*, vol. 249, pp. 1–12, 2013.

[42] X. Liu, D. Zhao, J. Zhou, W. Gao, and H. Sun, "Image interpolation via graph-based Bayesian label propagation," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1084–1096, 2014.

[43] X. Jia, L. Shang, B. Zhou, and Y. Yao, "Generalized attribute reduct in rough set theory," *Knowledge-Based Systems*, vol. 91, pp. 204–218, 2016.

[44] D. Chen and Y. Yang, "Attribute reduction for heterogeneous data based on the combination of classical and fuzzy rough set models," *IEEE Transactions on Fuzzy Systems*, vol. 22, no. 5, pp. 1325–1334, 2014.

[45] S. Xu, P. Wang, J. Li, X. Yang, and X. Chen, "Attribute reduction: an ensemble strategy," in *International Joint Conference on Rough Sets*, Lecture Notes in Computer Science, pp. 362–375, 2017.

[46] Y. Qian, J. Liang, W. Pedrycz, and C. Dang, "An efficient accelerator for attribute reduction from incomplete data in rough set framework," *Pattern Recognition*, vol. 44, no. 8, pp. 1658–1670, 2011.

[47] C. Wang, M. Shao, Q. He, Y. Qian, and Y. Qi, "Feature subset selection based on fuzzy neighborhood rough sets," *Knowledge-Based Systems*, vol. 111, pp. 173–179, 2016.

[48] X. Yang and Y. Yao, "Ensemble selector for attribute reduction," *Applied Soft Computing*, vol. 70, pp. 1–11, 2018.

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

[50] Y. Y. Yao and Y. Zhao, "Attribute reduction in decision-theoretic rough set models," *Information Sciences*, vol. 178, no. 17, pp. 3356–3373, 2008.

[51] A. Skowron and C. Rauszer, "The discernibility matrices and functions in information systems," *Fundamenta Informaticae*, vol. 11, pp. 331–362, 1992.

[52] F. Min, H. He, Y. Qian, and W. Zhu, "Test-cost-sensitive attribute reduction," *Information Sciences*, vol. 181, no. 22, pp. 4928–4942, 2011.

[53] J.-H. Dai and Y.-X. Li, "Heuristic genetic algorithm for minimal reduct in decision system based on rough set theory," in *Proceedings of 2002 International Conference on Machine Learning and Cybernetics*, pp. 833–836, China, November 2002.

[54] Y. Yao, Y. Zhao, and J. Wang, "On reduct construction algorithms," *Rough Sets & Knowledge Technology Proceedings*, vol. 4062, pp. 100–117, 2006.

[55] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[56] G. Xu, A. M. Hamouda, and B. C. Khoo, "Time-domain simulation of second-order irregular wave diffraction based on a hybrid water wave radiation condition," *Applied Mathematical Modelling: Simulation and Computation for Engineering and Environmental Systems*, vol. 40, no. 7-8, pp. 4451–4467, 2016.

[57] P. Wang, J. Wu, and X. Yang, "An improved method for computing eigenpair derivatives of damped system," *Mathematical Problems in Engineering*, vol. 2018, Article ID 8050132, 8 pages, 2018.

[58] W. Wei, B. Shaoyi, Z. Lanchun, W. Yongzhi, and Y. Hui, "Pavement roughness identification research in time domain based on neural network," *Journal of Vibroengineering*, vol. 17, no. 7, pp. 3865–3875, 2015.