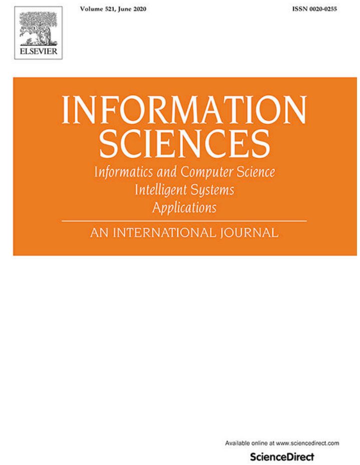
Journal Pre-proofs




# Three-way decision with co-training for partially labeled data

Can Gao, Jie Zhou, Duoqian Miao, Jiajun Wen, Xiaodong Yue

PII: S0020-0255(20)30870-7
DOI: https://doi.org/10.1016/j.ins.2020.08.104
Reference: INS 15819

To appear in: *Information Sciences*

Received Date: 23 April 2020
Revised Date: 16 July 2020
Accepted Date: 27 August 2020

Please cite this article as: C. Gao, J. Zhou, D. Miao, J. Wen, X. Yue, Three-way decision with co-training for partially labeled data, *Information Sciences* (2020), doi: https://doi.org/10.1016/j.ins.2020.08.104

# Three-way decision with co-training for partially labeled data

Can Gao[a,b], Jie Zhou[a,b*], Duoqian Miao[c], Jiajun Wen[a,b], Xiaodong Yue[d]

[a]College of Computer Science and Software Engineering, Shenzhen University
Shenzhen 518060, P.R. China
[b]SZU Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society
Shenzhen 518060, P.R. China
[c]Department of Computer Science and Technology, Tongji University
Shanghai 201804, China
[d]School of Computer Engineering and Science, Shanghai University
Shanghai 200444, China

Abstract

The theory of three-way decision plays an important role in decision making and knowledge reasoning. However, little attention has been paid to the problem of learning from partially labeled data with three-way decision. In this paper, we propose a three-way co-decision model for partially labeled data. More specifically, the problem of attribute reduction for partially labeled data is first investigated, and two semi-supervised attribute reduction algorithms based on novel confidence discernibility matrix are proposed. Then, a three-way co-decision model is introduced to classify unlabeled data into useful, useless, and uncertain data, and the model is iteratively retrained on the carefully selected useful data to improve its performance. Moreover, we theoretically analyze the effectiveness of the proposed model. The experimental results conducted on UCI data sets demonstrate that the proposed model is promising, and even compares favourably with the single supervised classifier trained on all training data with true labels.

Keywords: Three-way decision, semi-supervised reduct, confidence discernibility matrix, co-decision, partially labeled data

## 1. Introduction

The theory of rough sets [23] is an effective tool for handling vague, uncertain, or imprecise data. Since the pioneering work of Pawlak [22], several extended and generalized models have been proposed, such as neighbourhood rough sets [10, 45], covering rough sets [15, 42], fuzzy rough sets [4, 6], probabilistic rough sets [31, 32], and others [46]. Among them, three-way decision [30], proposed by Yao [33, 41], is one of the most popular and

*Corresponding author at: College of Computer Science and Software Engineering, shenzhen University, Shenzhen, 518060, P.R. China.
Email addresses: jie_jpu@163.com.

efficient models for decision-making. Despite originating from probabilistic rough sets [33], the research and development of three-way decision have gone beyond the realm of rough sets and become the methodology and philosophy for thinking in threes [35, 36, 38, 40, 47]. Due to the universality and effectiveness, three-way decision has been introduced to many research domains, such as attribute reduction [43], conflict analysis [37], formal concept analysis [39], etc. Decision-theoretic rough sets (referred to as DTRS hereafter) [42], as a representative paradigm of three-way decision, generalizes the Pawlak rough sets by introducing the theory of Bayesian risk decision. In DTRS, binary decisions with options "yes" and "no" are extended into triple decisions, i.e., "yes", "no", and "wait-to-see". Moreover, DTRS provides a unified and comprehensive framework for rough sets and exhibits the salient characteristics and advantages in probabilistic reasoning and semantic interpretation [34].

Both DTRS and other extensions of rough sets are primarily used to handle either labeled data or unlabeled data. However, in many real-world applications, such as web-page categorization, image retrieval, and intrusion detection [50], we often confront the case where labeled data are scarce since hand-labeled objects are fairly expensive to obtain, whereas unlabeled data are relatively cheap and readily available. In this scenario, traditional supervised learning may yield undesirable results because of the scarcity of labeled data, while unsupervised learning using only unlabeled data will result in the waste of valuable label information. Intuitively, a promising way is to fully capitalize on both labeled and unlabeled data to train an effective learning model [48, 50].

For the data containing both labeled and unlabeled data (referred to as partially labeled data hereafter), Lingras [14] et al. extended DTRS from two classes to multiple classes and introduced semi-supervised costs for promotional campaigns in real-world retail stores. Miao et al. [18] developed a semi-supervised discernibility matrix and proposed a diverse semi-supervised reducts-based model for partially labeled data. Dai et al. [5] employed the consistent rate of objects as the fitness function to generate semi-supervised reduct. Based on the concept of discernibility, Dai et al. [7] further developed two attribute reduction measures for partially labeled data. Instead of equivalence relation, fuzzy or neighbourhood relations-based rough set models are also introduced to deal with partially labeled data. Parthalain and Jensen [21] employed the unlabeled objects contained in the fuzzy lower approximation of all decision classes to retrain the model iteratively and presented a fuzzy rough set-based self-training model for partially labeled data. Wang et al. [29] used Gaussian kernel-based fuzzy rough set to measure the inconsistency of unlabeled objects and proposed a SVM-based sample selection algorithm for active learning. Jensen et al. [11] presented a semi-supervised fuzzy rough attribute reduction method, in which the fuzzy dependency degree on both labeled and unlabeled data was used to measure the quality of attribute subsets. To deal with numerical data, Liu et al. [16] introduced a weighted neighbourhood approximate quality and neighbourhood granules for partially labeled data. Further, they [17] used a graph-based semi-supervised method to yield the pseudo labels of all unlabeled data, and local neighbourhood decision error rates under different decision classes were combined to measure the significance of attributes. Li et al. [13] provided a semi-supervised attribute reduction method for partially labeled data with numerical attributes, where conditional neighbourhood granulation and neighbourhood granulation were used to measure the significance of attributes on labeled data and unlabeled data, respectively. By integrating cost-sensitive learning and three-way theory, Min et al. [19] proposed an active learning algorithm for classification. Qian et al. [24, 25] presented several local rough set models for big data with limited labels and provided some efficient local attribute reduct algorithms based on local lower approximation. In addition, the theory of rough sets has also been successfully applied to practical problems with partially labeled data [12, 26].

The aforementioned works mainly concentrate on rough sets-based semi-supervised attribute reduction or practical applications. Little attention has been paid for the semi-supervised rough set model to learn directly from both labeled and unlabeled data. On the one hand, the utilization of unlabeled data is a key problem of semi-supervised learning model, and unlabeled data may contain noisy or useless objects, which have a negative effect on the learning model. To guarantee the performance of semi-supervised learning model, it is vital and necessary to develop an appropriate and effective mechanism to select useful unlabeled objects. On the other hand, decision-making under uncertainty often results in different costs or risks. The selection of unlabeled objects should take into consideration the cost or risk of decision. Motivated by the above facts, we propose a three-way decision-based semi-supervised model for partially labeled data. The main contribution of this paper is threefold.

(1) To address the problem of attribute reduction for partially labeled data, we develop the concept of confidence discernibility matrix, based on which a heuristic algorithm is designed to yield the optimal reduct of partially labeled data. The confidence discernibility matrix takes into consideration both labeled and unlabeled data and allows a certain degree of inconsistency, thus resulting in better adaptability and robustness for partially labeled data. In addition, we prove several propositions about the confidence discernibility matrix, which provide the theoretical basis for semi-supervised attribute reduction.

(2) To exploit unlabeled data efficiently, we design a three-way co-decision model for partially labeled data. The unlabeled objects to use have a considerable effect on the performance of the learning mode. Three way-decision is an effective method for decision making under uncertainty and risk. We thus introduce the theory of three-way decision to conduct the selection of useful unlabeled data. Moreover, motivated by the idea of co-training [2], the collaborative decision framework using two distinct semi-supervised reducts is adopted, which could make the classifiers of the model learn from each other. By incorporating the theory of three-way decision with the mechanism of co-training, the co-decision model could make full use of unlabeled data to improve its performance.

(3) To gain a deep insight into the proposed model, we theoretically analyze the model from the perspective of noise learning and give the upper bound on the number of exploitable unlabeled data. Additionally, extensive experiments are performed to test the effectiveness of the proposed model, and promising

results are achieved, indicating the potential of the proposed model for partially labeled data.

The rest of this paper is organized as follows. Section 2 presents some concepts in semi-supervised learning and three-way decision, respectively. Section 3 describes the proposed co-decision model for partially labeled data, and its effectiveness is also theoretically analyzed. Experimental results and analysis are shown in Section 4. Finally, Section 5 concludes the paper and indicates future research work.

## 2. Preliminaries

This section will briefly review some concepts related to semi-supervised learning and three-way decision. More details about these theories can be found in [32-41, 50].

### 2.1. Semi-supervised learning

In semi-supervised learning, we are provided with a partially labeled data $U = L \cup N$ with $l + n$ objects described by m-dimensional attributes, where $l$ number of labeled objects $L = \{x_i, y_i\}_{i=1}^{l}$ are labeled and $n$ number of unlabeled objects $N = \{x_i, ?\}_{i=l+1}^{l+n}$ $(l \ll n)$ are unlabeled. In the context of semi-supervised learning, we can, on the one hand, use labeled data to enhance the quality of unsupervised clustering, called semi-supervised clustering [50]. On the other hand, unlabeled data can be utilized to improve the performance of the supervised models that learn only from labeled data, called semi-supervised classification or regression [49]. The detailed description of these methods could refer to [28, 50]. In this paper, we only focus on semi-supervised classification.

Semi-supervised classification aims at using a large amount of unlabeled data to aid the training of supervised models when labeled data at hand are scarce. Roughly speaking, semi-supervised classification can be further categorized into generative methods, low-density separation methods, graph-based methods, and disagreement-based methods [28]. Co-training [2, 3] is one of the most popular multi-view models and has been applied to many practical problems successfully. Standard co-training assumes that each object can be described by two sufficient and redundant attribute subsets (views). On each attribute subset, a base classifier is first trained on initial labeled data. By labeling the most confident unlabeled objects to their counterparts, the two base classifiers learn from each other iteratively and are retrained on their enlarged training sets to improve the performance.

Unfortunately, in practical applications, it is difficult to meet the assumption of two naturally partitioned attribute subsets in co-training. Although some compromise solutions have been proposed, such as random subspace, resampling, and heterogeneous algorithms [28], it is still an open question on how to split a natural attribute set into two attribute subsets. Furthermore, the performance of co-training is highly related to the quality of unlabeled data used in the learning process. In standard co-training, the highly confident objects are usually selected to enlarge the training sets of base classifiers, and the evaluation criteria for confident objects, such as classification accuracy, cross-validation, majority voting, and data editing, are often used [28]. However, these criteria do not consider the misclassification cost of unlabeled objects. It seems unreasonable when different decisions have different misclassification costs.

### 2.2. Three-way decision

The theory of three-way decision is a methodology for decision-making with the alternatives of acceptance, rejection, and noncommitment. Decision-theoretic rough sets (DTRS), as an extension of rough sets, is one of the most popular models in three-way decision and has witnessed a rapid growth of interest in theory and applications [32-39, 41]. In what follows, we will review some related concepts about DTRS.

In DTRS, the data to deal with is called an information system [23] and is denoted as $IS = (U,A,V,f)$, where $U$ is the set of objects, called the universe; $V$ is the set of attributes to describe the objects; $V$ is the union of attribute domains such that $V = \cup V_a$ for $a \in A$, where $V_a$ denotes the domain of the attribute $a$; and $f$ is an information function that associates each attribute of an object belonging to $U$ with a unique value such that $f(x,a) \in V_a$ for each $x \in U$ and $a \in A$. The information system is also called a decision information system or decision table if the attribute set $A$ can be further divided into the condition attribute set $C$ and the decision attribute set $D$ [23].

For an attribute subset $B$ of $A$, it partitions the universe $U$ into a family of equivalence classes $U/B$. An equivalence class containing $x$ is denoted as $[x]_B$ and is referred to as $B$-elementary set or $B$-elementary granule [23]. Let $X$ be a subset of the universe $U$, the lower approximation $\underline{B}(X)$ and the upper approximation $\overline{B}(X)$ with respect to $B$ are defined as [23]:

$$\begin{aligned} \underline{B}(X) &= \{x \in U \mid [x]_B \subseteq X\}, \\ \overline{B}(X) &= \{x \in U \mid [x]_B \cap X \neq \emptyset\}. \end{aligned} \tag{1}$$

The $B$-lower approximation of $X$ is also called the $B$-positive region $POS_B(X)$ of $X$ over $U$. The set-theoretic difference of the $B$-upper and $B$-lower approximations is called the $B$-boundary region $BND_B(X)$ of $X$ over $U$, i.e., $BND_B(X) = \overline{B}(X) - \underline{B}(X)$. The universe after removing the objects in the $B$-upper approximation is called the $B$-negative region $NEG_B(X)$ of $X$ over $U$, i.e., $NEG_B(X) = U - \overline{B}(X)$.

Let $U/D = \{Y_1, Y_2, \ldots, Y_{|U/D|}\}$ be the partition induced by the decision attribute $D$ over $U$. The positive, boundary, and negative regions of $D$ with respect to $C$ are defined as [23]:

$$\begin{aligned} POS_C(D) &= \bigcup_{Y_i \in U/D} \underline{C}(Y_i), \\ BND_C(D) &= \bigcup_{Y_i \in U/D} (\overline{C}(Y_i) - \underline{C}(Y_i)), \\ NEG_C(D) &= U - \bigcup_{Y_i \in U/D} \overline{C}(Y_i). \end{aligned} \tag{2}$$

Let $\Omega = \{X, X^C\}$ be a set of states indicating an object $x$ is in $X$ or not in X, respectively, and $\Lambda = \{a_P, a_B, a_N\}$ be a set of actions deciding the object $x$ to be

$POS(X)$, $BND(X)$, or $NEG(X)$, respectively. The cost functions taking different actions under the states $X$ and $X^C$ can be expressed as Table 1 [33]:

Table 1: Cost functions for different actions under the states X and $X^C$.

| | $a_P$ | $a_B$ | $a_N$ |
|---|---|---|---|
| $X$ | $\lambda_{PP}$ | $\lambda_{BP}$ | $\lambda_{NP}$ |
| $X^C$ | $\lambda_{PN}$ | $\lambda_{BN}$ | $\lambda_{NN}$ |

In the table, $\lambda_{PP}$,$\lambda_{BP}$, and $\lambda_{NP}$ denote the costs caused by taking the actions $a_P$,$a_B$ and $a_N$, respectively, when the object $x$ belongs to $X$, and $\lambda_{PN}$, $\lambda_{BN}$, and $\lambda_{NN}$ denote the costs caused by taking the same actions but the object x does not belong to $X$.

Given an object $x$, the expected costs of taking different actions can be defined as [33]:

$$\begin{aligned} R(a_P|[x]) &= \lambda_{PP}P(X|[x]) + \lambda_{PN}P(X^C|[x]), \\ R(a_B|[x]) &= \lambda_{BP}P(X|[x]) + \lambda_{BN}P(X^C|[x]), \\ R(a_N|[x]) &= \lambda_{NP}P(X|[x]) + \lambda_{NN}P(X^C|[x]), \end{aligned} \tag{3}$$

where $P(X|[x])$ and $P(X^C|[x])$ denote the probabilities that the object $x$ belongs to X and $X^C$, respectively, and $P(X|[x]) = 1 - P(X^C|[x])$.

According to Bayesian decision theory, the following minimum-risk rules can be derived [33]:

(P) if $R(a_P|[x]) \leq min\{R(a_B|[x]),\ R(a_N|[x])\}$, then decide $x \in POS(X)$;
(B) if $R(a_B|[x]) \leq min\{R(a_P|[x]),\ R(a_N|[x])\}$, then decide $x \in BND(X)$;
(N) if $R(a_N|[x]) \leq min\{R(a_P|[x]),\ R(a_B|[x])\}$, then decide $x \in NEG(X)$.

Suppose the inequality $(\lambda_{PN} - \lambda_{BN})(\lambda_{NP} - \lambda_{BP}) > (\lambda_{BP} - \lambda_{PP})(\lambda_{BN} - \lambda_{NN})$ holds, the decision rules can be further simplified as [33]:

(P) if $P(X|[x]) \geq \alpha$, then decide $x \in POS(X)$;
(B) if $\beta < P(X|[x]) < \alpha$, then decide $x \in BND(X)$;
(N) if $P(X|[x]) \leq \beta$, then decide $x \in NEG(X)$,

where

$$\begin{aligned} \alpha &= \frac{\lambda_{PN} - \lambda_{BN}}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})}, \\ \beta &= \frac{\lambda_{BN} - \lambda_{NN}}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})}. \end{aligned} \tag{4}$$

Given the parameters $\alpha$ and $\beta$, the lower and upper approximations can be redefined as in [33]:

$$\begin{aligned} \underline{B}_{(\alpha,\beta)}(X) &= \{x \in U \mid \mu_B(x) \geq \alpha\}, \\ \overline{B}_{(\alpha,\beta)}(X) &= \{x \in U \mid \mu_B(x) > \beta\}. \end{aligned} \tag{5}$$

Similarly, the positive, boundary, and negative regions can be defined as [33]:

$$\begin{aligned} POS_C^{(\alpha,\beta)}(D) &= \{x \in U | P(D_{max}([x]_C) | [x]_C) \geq \alpha\}, \\ BND_C^{(\alpha,\beta)}(D) &= \{x \in U | \beta < P(D_{max}([x]_C) | [x]_C) < \alpha\}, \\ NEG_C^{(\alpha,\beta)}(D) &= \{x \in U | P(D_{max}([x]_C) | [x]_C) \leq \beta\}, \end{aligned} \tag{6}$$

where $D_{max}([x]_C) = argmax_{D_i \in U/D}\{P(D_i | [x]_C)\}$.

## 3. Three-way decision-based co-decision model for partially labeled data

In this section, we first describe the overall framework of the proposed model. The concept of confidence discernibly matrix is then provided and used to yield the reducts of partially labeled data. Subsequently, a three-way decision-based co-decision model is presented based on two distinct semi-supervised reducts. Finally, the model is theoretically analyzed.

### 3.1. Overall framework of the proposed model

Traditional models in three-way decision mainly deal with labeled or unlabeled data, and one classifier is often used in the learning process. Due to the scarcity of labeled objects, learning models with only one classifier may be insufficient and undesired. In fact, some data sets, especially when there are a large number of attributes, usually have more than one reduct, and each reduct could describe the data completely and competently. Additionally, these reducts reflect the data from different viewpoints, thus resulting in different inductive biases. Intuitively, one could take advantage of the diversity of multiple reduct subspaces to construct an efficient multi-view model for partially labeled data. Bearing this in mind, we propose a distinct reduct subspaces-based co-decision model for partially labeled data (see Figure 1).

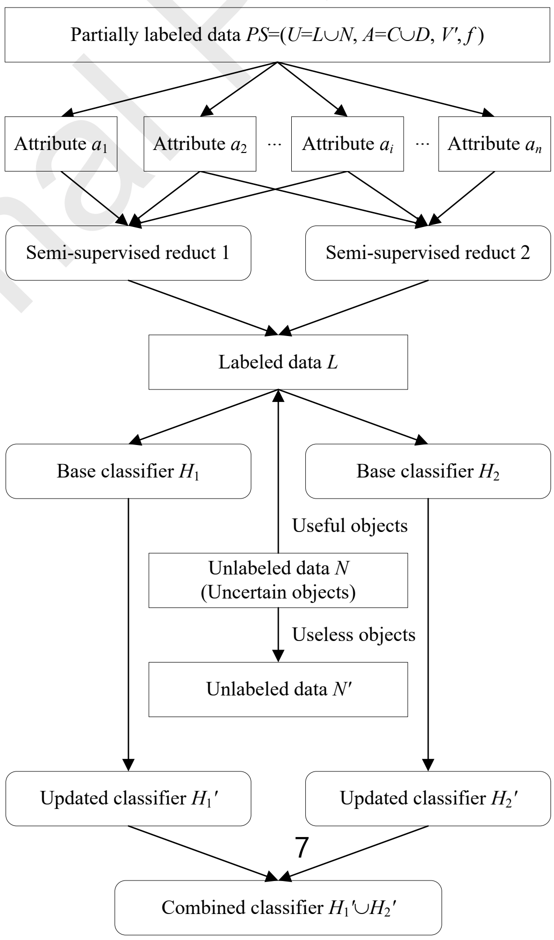


Figure 1. Framework of three-way decision-based co-decision model for partially labeled data

More specifically, a semi-supervised attribute reduction algorithm is first used to generate two distinct reducts of partially labeled data, on each of which a base classifier is trained with initial labeled data. Then two base classifiers learn from each other iteratively by tagging some useful unlabeled objects with minimum risks to their companions until there is no eligible unlabeled object. After improved on unlabeled data, the two classifiers are combined to form the final classifier. In the following sections, we will elaborate on the proposed model.

### 3.2. Confidence discernibility matrix-based attribute reduction for partially labeled data

Attribute reduction (feature selection) [8, 43] is a process of removing irrelevant and redundant attributes from data and has become an important pre-processing step in machine learning and pattern recognition. It could not only speed up the learning process, but also weaken the problem of over-fitting. Attribute reduction is one of the most important applications of rough sets, and several attribute reduction methods have been proposed [46]. Among them, the methods based on the discernibility matrix [27, 44] are commonly used and has attracted much attention due to its simplicity and monotonicity. Formally, the discernibility matrix and its reduct can be defined as follows.

Definition 1. Let $IS = (U, A = C \cup D, V, f)$ be a decision table. The element of the discernibility matrix M is denoted as [27]:

$$e_{ij} = \begin{cases} \{a \in C | a(x_i) \neq a(x_j)\}, & d(x_i) \neq d(x_j) \\ \emptyset, & otherwise \end{cases} \quad (7)$$

Definition 2. Let $IS = (U, A = C \cup D, V, f)$ be a decision table and M be the discernibility matrix of $IS$. An attribute $a \in C$ is a core attribute if and only if there exists a singleton $e$ in M such that $e = \{a\}$ [44].

Definition 3. Let $IS = (U, A = C \cup D, V, f)$ be a decision table and M be the discernibility matrix of IS. For an attribute subset P of C, P is a reduct of C if and only if [44]:

(I) $\forall e \in M \wedge e \neq \emptyset, P \cap e \neq \emptyset$, and

(II) $\forall a \in P \wedge P^* = P - \{a\}, \exists e \in M \wedge P^* \cap e = \emptyset.$

According to the definition, a reduct is a subset of condition attributes that has an intersection with any non-empty element in the discernibility matrix. Existing discernibility matrix-based methods mainly deal with labeled or unlabeled data. However, partially labeled data comes with both labeled and unlabeled data. To address this problem, a new discernibility matrix is developed to handle partially labeled data.

Generally, a partially labeled data consists of few labeled objects but plenty of unlabeled objects. Intuitively, a reduct of partially labeled data should be able to distinguish both labeled and unlabeled objects. Therefore, in the process of attribute reduction, it is desired that the method of attribute reduction could take into consideration all kinds of objects. Moreover, in partially labeled data, the initial labeled data may be noisy, and the unlabeled data to be used is full of uncertainty so that a probabilistic method of attribute reduction is preferred. To this end, we propose a novel concept of confidence discernibility matrix, which takes into consideration the discernible information and probability distribution of both labeled and unlabeled objects. In what follows, we will give an example to illustrate the proposed discernibility matrix.

Example 1. Let $PS = (U = L \cup N, A = C \cup D, V', f)$ be a partially labeled data shown in Table 2, where $U = \{x_1, x_2, ..., x_{15}\}$, $C = \{a_1, a_2, ..., a_7\}$, $V_a = \{0,1\}$ for every $a \in C$, and $V_D = \{d_1, d_2, d_3, ?\}$.

Table 2: A partially labeled data

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | $d$ |
|---|---|---|---|---|---|---|---|---|
| $x_1$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | $d_1$ |
| $x_2$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | $d_1$ |
| $x_3$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | $d_2$ |
| $x_4$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | $d_2$ |
| $x_5$ | 0 | 0 | 0 | 0 | 0 | 1 | 1 | $d_2$ |
| $x_6$ | 0 | 0 | 0 | 0 | 0 | 1 | 1 | $d_3$ |
| $x_7$ | 0 | 0 | 0 | 0 | 0 | 1 | 1 | $d_3$ |
| $x_8$ | 0 | 0 | 0 | 0 | 1 | 1 | 1 | $d_1$ |
| $x_9$ | 0 | 0 | 0 | 0 | 1 | 1 | 1 | $d_3$ |
| $x_{10}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ? |
| $x_{11}$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | ? |
| $x_{12}$ | 0 | 0 | 0 | 1 | 1 | 1 | 1 | ? |
| $x_{13}$ | 0 | 0 | 0 | 1 | 1 | 1 | 1 | ? |
| $x_{14}$ | 1 | 0 | 1 | 0 | 1 | 1 | 1 | ? |
| $x_{15}$ | 1 | 0 | 1 | 0 | 1 | 1 | 1 | ? |

In the table, under all condition attributes, the universe is partitioned into six condition equivalence classes, i.e., $U/C = \{\{x_1,x_{10}\},\{x_2,x_3,x_4,x_{11}\},\{x_5, x_6,x_7\},\{x_8,x_9\},\{x_{12},x_{13}\},\{x_{14},x_{15}\}\}$. For the equivalence class $\{x_1,x_{10}\}$, there are two objects, one of which is labeled and the other is unlabeled. Undoubtedly, the class information of the labeled object can be propagated to the unlabeled one because the two objects have the same description in each condition attribute. The decision of the object $x_{10}$ can thus be changed from "?" to $d_1$. The equivalence class $\{x_2,x_3,x_4,x_{11}\}$ consists of labeled and unlabeled objects with different kinds of decisions, i.e., the real decisions $d_1$ and $d_2$, and the unknown decision "?". Actually, this equivalence class is inconsistent. In this case, the majority decision of all labeled objects in the equivalence class can be assigned to the unlabeled objects. Therefore, the unlabeled object $x_{11}$ can be labeled the real decision $d_2$. For other unlabeled objects $x_{12},x_{13},x_{14}$, and $x_{15}$, we consider them as the objects with a special pseudo decision "*", whose decisions will be replaced by the real decisions during the learning process. Finally, each object in the table has a real decision or a pseudo decision, and the partially labeled data becomes a pseudo decision table. To deal with the pseudo decision table transformed from partially labeled data, we introduce a new confidence discernibility matrix.

Definition 4. Let $PS = (U = L \cup N, A = C \cup D, V', f)$ be a partially labeled data. For an object $x \in U$, its maximum inclusion degree and majority decision are denoted as $MP(x) = max\{P(D_1|[x]_C), P(D_2|[x]_C), \ldots, P(D_{|U/D|}|[x]_C)\}$ and $MD(x) = argmax_{D_i \in U/D}\{P(D_i|[x]_C)\}$, respectively.

Definition 5. Let $PS = (U = L \cup N, A = C \cup D, V', f)$ be a partially labeled data and $\delta$ be a confidence threshold parameter. The element of the confidence discernibility matrix $CM(\delta)$ of $PS$ is denoted as:

$$e_{ij}(\delta) = \begin{cases} \{a \in C | a(x_i) \neq a(x_j)\}, & \begin{array}{c} (MD(x_i) \neq MD(x_j) \\ \vee MD(x_i) = * \vee MD(x_j) = * ) \\ \wedge (max\{MP(x_i), MP(x_j)\} \geq \delta) \end{array} \\ \emptyset, & otherwise \end{cases} \tag{8}$$

When the maximum inclusion degree of an object is lower than 1, there are different definitions for the element of the discernibility matrix and may generate different discernible information. In existing discernibility matrix [27], the discernible information of all inconsistent objects is either all retained or discarded. In fact, the measure of classification ability under uncertainty is reflected not only in the decision, but also in the maximum confidence the inductive decision rule has. In Definition 5, besides the decision information, a confidence threshold parameter is introduced to determine the discriminating information of the discernibility matrix. As a result, the discernible information is generated only when two objects have different majority decisions and at least one of the two objects has a maximum inclusion degree greater than $\delta$. Compared to traditional discernibility matrices, the proposed confidence discernibility matrix ignores the information generated by each pair of objects whose maximum inclusion degree is all less than $\delta$. Actually, in the case of decision-making with uncertainty, that kind of information, in a sense, is not necessary for classification and may increase the complexity of attribute reduction. Therefore, we should remove them to make the discernibility matrix more concise and efficient.

Formally, the pseudo decision table transformed from partially labeled data $PS = (U = L \cup N, A = C \cup D, V', f)$ is denoted as $TS = (U', A = C \cup D, V', f)$, while the decision table after labeling all unlabeled objects in the $PS$ with ground-truth decisions is denoted as $IS = (U, A = C \cup D, V, f)$ (called the ground-truth decision table). In what follows, we will discuss the properties of the proposed confidence discernibility matrix.

Proposition 1. Let $PS = (U = L \cup N, A = C \cup D, V', f)$ be a partially labeled data and $\delta$ be a confidence threshold parameter. If $CM_1^\delta$ is the confidence discernibility matrix of the ground-truth decision table $IS$, and $CM_2^\delta$ is the confidence discernibility matrix of the transformed decision table $TS$, then, for each element $e_{ij}^1 \in CM_1^\delta$, there is $e_{ij}^2 \in CM_2^\delta$ such that $e_{ij}^1 \subseteq e_{ij}^2$.

Proof. Without loss of generality, assume $x_i$ and $x_j$ are two objects in the partially labeled data PS. In terms of their decision values, there are three different cases, i.e., $x_i \in L \wedge x_j \in L$, $x_i \in L \wedge x_j \in N$ or $x_i \in N \wedge x_j \in L$, and $x_i \in N \wedge x_j \in N$.

(1) Case 1: $x_i \in L \wedge x_j \in L$. Since the two objects $x_i$ and $x_j$ are all labeled, there is no difference between the elements $e_{ij}^1$ and $e_{ij}^2$, i.e., $e_{ij}^1 = e_{ij}^2$.

(2) Case 2: $x_i \in L \wedge x_j \in N$ or $x_i \in N \wedge x_j \in L$. In this case, only one object is labeled. But each unlabeled object of $PS$ is assigned a certain decision or a pseudo decision "*" after transformation. Thus, $e_{ij}^2$ may be a non-empty element in $CM_2^\delta$. While, in the ground-truth decision table $IS$, all objects have certain decisions, and the element $e_{ij}^1$ is an empty set when the objects $x_i$ and $x_j$ have the same decision. Thus, the element $e_{ij}^1$ of $CM_1^\delta$ is a subset of the element $e_{ij}^2$ of $CM_2^\delta$, i.e., $e_{ij}^1 \subseteq e_{ij}^2$.

(3) Case 3: $x_i \in N \wedge x_j \in N$. Since both objects are unlabeled, the element $e_{ij}^2$ of $CM_2^\delta$ is definitely non-empty when the objects $x_i$ and $x_j$ have distinct values in their condition attributes. However, in the ground-truth decision table $IS$, the two objects may have the same decision so that the element $e_{ij}^1$ may be an empty set. Thus, $e_{ij}^1 \subseteq e_{ij}^2$.

Therefore, in every possible case, we have $e_{ij}^1 \subseteq e_{ij}^2$. The proposition is proved.

Proposition 2. Let $PS = (U = L \cup N, A = C \cup D, V', f)$ be a partially labeled data and $\delta$ be a confidence threshold parameter. If $Core_1$ is the set of core attributes in the ground-truth decision table $IS$, and $Core_2$ is the set of core attributes in the transformed decision table $TS$, then $Core_1 \subseteq Core_2$.

Proof. According to Definition 2 and Proposition 1, it is straightforward to draw the conclusion.

Proposition 3. Let $PS = (U = L \cup N, A = C \cup D, V', f)$ be a partially labeled data and $\delta$ be a confidence threshold parameter. If $Red_1$ is a reduct of the ground-truth decision table $IS$, then there must exist a reduct $Red_2$ of the transformed decision table $TS$ such that $Red_1 \subseteq Red_2$.

Proof. Assume that $CM_1^\delta$ and $CM_2^\delta$ are the confidence discernibility matrices of the ground-truth decision table $IS$ and the transformed decision table $TS$, respectively, and $Red_1$ is a reduct in $CM_1^\delta$. Without loss of generality, assume the difference set between the elements of $CM_1^\delta$ and $CM_2^\delta$ has only one element $e$, i.e., $CM_2^\delta = CM_1^\delta \cup e$. We proceed by cases:

(1) Case 1: $\exists e' \in CM_1^\delta \wedge e' \subseteq e$. According to the definition for attribute reduction (see Definition 3), each non-empty element in $CM_1^\delta$ has a non-empty intersection with the reduct so that $Red_1 \cap e' \neq \emptyset$. Since $e' \subseteq e$, we have $Red_1 \cap e \neq \emptyset$. Thus, $Red_1$ is also a reduct in $CM_2^\delta$, and $Red_1 = Red_2$.

(2) Case 2: $\exists e' \in CM_1^\delta \wedge e' \supset e$. Since $e' \supset e$, the reduct $Red_1$ may have a non-empty intersection with $e' - e$ so that $Red_1 \cap e = \emptyset$. However, the reduct $Red_1$ after adding an attribute $a \in e$ can be a reduct $Red_2$ in $CM_2^\delta$. Thus, we have $Red_1 \subseteq Red_2$.

(3) Case 3: $\forall e' \in CM_1^{\delta} \wedge (e' \not\subset e \wedge e' \not\supset e)$. Since the element $e$ neither contains nor be contained by any element of $CM_1^{\delta}$, the reduct $Red_1$ in $CM_1^{\delta}$ may not be a reduct in $CM_2^{\delta}$. But there exists at least one attribute $a \in e$ such that $Red_2 = Red_1 \cup \{a\}$ is a reduct in $CM_2^{\delta}$. Thus, we have $Red_1 \subseteq Red_2$.

Thus, in every possible case, we have $Red_1 \subseteq Red_2$. The proposition is proved.

The above propositions indicate that, for any possible ground-truth decision table derived from partially labeled data, there definitely exists a reduct in the transformed decision table such that the reduct has the full ability to discern all objects in the ground-truth decision table. On the basis of this fact, we can investigate the problem of attribute reduction for partially labeled data on the transformed decision table.

Definition 6. Let $PS = (U = L \cup N, A = C \cup D, V', f)$ be a partially labeled data and $CM(\delta)$ be the confidence discernibility matrix of the transformed decision table of PS under the confidence threshold $\delta$. Then, for any condition attribute $a \in C$, its relevant set is defined as:

$$RM_{\mathrm{CM}}^{\delta}(a) = \{e \in CM(\delta) | a \in e\}. \tag{9}$$

Definition 7. Let $PS = (U = L \cup N, A = C \cup D, V', f)$ be a partially labeled data and $CM(\delta)$ be the confidence discernibility matrix of the transformed decision table of PS under the confidence threshold $\delta$. Then, for any condition attribute $a \in C$, the complement set with respect to its relevant set is defined as:

$$OM_{\mathrm{CM}}^{\delta}(a) = \{e - \{a\} | e \in RM_{\mathrm{CM}}^{\delta}(a)\}. \tag{10}$$

In the definitions, the relevant set of an attribute consists of the elements that contain the attribute. While, in the relevant set, the elements after deleting the attribute itself constitute the complement set of the attribute.

On the basis of the set operators defined above, an attribute reduction algorithm can be developed to obtain the reduct of partially labeled data. However, finding the minimal reduct of a given data is NP-hard so that heuristic algorithms are preferred. In practice, due to high efficiency and effectiveness, the forward search strategy by iteratively adding attributes is often used. In this paper, we also adopt the forward search strategy to maximize the discernibility ability of the selected attributes with respect to the confidence discernibility matrix. The procedure can be depicted by Algorithm 1.

In the algorithm, the partially labeled data is first transformed into a pseudo decision table, and the confidence discernibility matrix is computed under the confidence threshold parameter (line 1 and line 2). After putting the singletons into the reduct, the algorithm iteratively selects the optimal attributes into the reduct and simultaneously removes their relevant sets until the confidence discernibility matrix is empty (line 3 to line 8). The optimal semi-supervised reduct is finally generated after the algorithm terminates, which has a non-empty intersection with any non-empty element of the confidence discernibility matrix, thus preserving the same discriminating power as all condition attributes.

---

Algorithm 1 A confidence discernibility matrix-based semi-supervised attribute reduction algorithm for partially labeled data

---

Input:

A partially labeled data $PS = (U = L \cup N, A = C \cup D, V', f)$ and a confidence threshold parameter $\delta$;

Output:

An optimal semi-supervised reduct $P$;

1: Transform the partially labeled data $PS$ into a pseudo decision table $TS$;
2: Compute the confidence discernibility matrix $CM(\delta)$ of $TS$, $P \leftarrow \emptyset$;
3: Add all singletons of $CM(\delta)$ into $P$ and remove their relevant sets from $CM(\delta)$;
4: While $CM(\delta) \neq \emptyset$ Do
5: Select an attribute $a_{opt}$ that has the maximum frequency within $CM(\delta)$;
6: $P \leftarrow P \cup \{a_{opt}\}$;
7: $CM(\delta) \leftarrow CM(\delta) - RM_{\mathrm{CM}}^{\delta}(a_{opt})$ //Remove the relevant set of $a_{opt}$;
8: End While
9: Return The semi-supervised reduct $P$.

---

Without loss of generality, assume that a partially labeled data has $|U|$ objects described by $|C|$ attributes. The time cost for constructing a confidence discernibility matrix is $O(|C||U|^2)$. In each iteration, the algorithm selects an optimal attribute and simultaneously removes the relevant set from the confidence discernibility matrix. In the worst-case, the matrix is empty after $|C|$ rounds of selection. Therefore, based on the confidence discernibility matrix, the time cost for computing an optimal reduct is $O(|C|^2|U|^2)$. The total time cost of Algorithm 1 is $O(|C||U|^2) + O(|C|^2|U|^2)$, which is approximate to $O(|C|^2|U|^2)$, and the total space cost is at most $O(|C||U|^2)$.

### 3.3. Co-decision model for partially labeled data

In traditional three-way decision-based classification, learning model mainly deals with labeled data and trains only one classifier. However, a partially labeled data usually contains few labeled data but along with a large amount of unlabeled data. Obviously, due to the scarcity of labeled data, the learning model with one classifier is not sufficient. Co-training is a multi-view paradigm that has been proved to be effective for partially labeled data [2]. It trains two classifiers on initial labeled data and achieves better performance by learning from unlabeled data. Standard co-training relies heavily on two sufficient and redundant subsets of attributes to train its classifiers. However, most real-world data have only one undivided set of attributes. In order to use the paradigm of co-training, we need to address the problem of splitting the whole attribute set into two attribute subsets.

Based on Algorithm 1, we can obtain an optimal reduct of partially labeled data. It can be one attribute subset for co-training because each reduct is a jointly sufficient subset of attributes to discriminate all objects in partially labeled data. As for the other attribute subset, the theoretically best way is to obtain all reducts of partially labeled data and then select the reduct that has the least common attributes with the optimal reduct. However, finding all reducts is very time-consuming, and thus the heuristic algorithm is preferred. Based on the concept of the complement set of an attribute (see Definition 7), we can develop a heuristic algorithm to yield another distinct reduct by slightly adjusting the procedure of Algorithm 1. More specifically, in each round of attribute selection, Algorithm 1 will select an optimal attribute and discard the relevant set of the optimal attribute. According to Definitions 6 and 7, the relevant set of an attribute consists of the attribute itself and its complement set. In fact, the attributes in the complement set can also be used to yield the reducts. Therefore, we can use the redundancy of attributes to generate two distinct reducts. The procedure is shown in Algorithm 2.

---

Algorithm 2 A heuristic algorithm for distinct semi-supervised reducts

---

Input:
A partially labeled data $PS=(U=L\cup N, A=C\cup D, V', f)$ and a confidence threshold parameter $\delta$;
Output:
Two distinct semi-supervised reducts $P_1$ and $P_2$;

```
Transform the partially labeled data PS into a pseudo decision table TS;
Compute the confidence discernibility matrix CM(δ) of TS, Core←∅;
Add all singletons in CM(δ) to Core and remove their relevant sets from
  CM(δ), P_1←Core, P_2←Core , CM_1^δ←CM(δ), CM_2^δ←∅;
While CM_1^δ ≠ ∅ Do
    Select an attribute a_opt that has the maximum frequency within CM_1^δ;
    P_1←P_1 ∪ {a_opt} and CM_1^δ←CM_1^δ − RM_{CM_1}^δ(a_opt);
    CM_2^δ←CM_2^δ ∪ OM_{CM_1}^δ(a_opt); //Information for another reduct
End While
Add all singletons of CM_2^δ into P_2 and remove their relevant sets from
  CM_2^δ;
While CM_2^δ ≠ ∅ Do
    Select an attribute a_opt that has the maximum frequency within CM_2^δ;
    P_2←P_2 ∪ {a_opt} and CM_2^δ←CM_2^δ − RM_{CM_2}^δ(a_opt);
End While
Return Two semi-supervised reducts P_1 and P_2.
```

---

In Algorithm 2, after computing the confidence discernibility matrix of the partially labeled data, the core attributes, that is, the attributes in the singletons of the confidence discernibility matrix, are first added into each semi-supervised reduct, and their relevant sets are removed accordingly. The algorithm, on the one hand, iteratively selects the optimal attributes from the current confidence discernibility matrix to form the optimal reduct. On the other hand, the complement sets of the selected optimal attributes are reserved for the other reduct. The elements after removing an attribute may become the singletons so that all singletons in the collection of the complement sets are first added into the second reduct. The algorithm repeatedly selects the optimal attributes in the current collection of the complement sets until the collection is empty. Since the second reduct is generated from the complement sets of all selected optimal attributes in the optimal reduct, the two reducts will be different and diverse. For Table 2, the confidence discernibility matrix after the law of absorption is $\{\{a_6\},\{a_7\},\{a_5\},\{a_4\},\{a_1,a_3\}$, and two reducts $\{a_6,a_7,\ a_5,a_4,a_1\}$ and $\{a_6,a_7,a_5,a_4,a_3\}$ can be generated by Algorithm 2.

As for the complexity, Algorithm 2 performs the process of Algorithm 1 twice, thus its time and space cost is almost the same as that of Algorithm 1, i.e., $O(|C|^2|U|^2)$ and $O(|C||U|^2)$.

To efficiently learn from partially labeled data, we also need to address the problem of selecting unlabeled objects because not all unlabeled data is beneficial to the learning model. Generally, unlabeled data can be divided into useful, useless, and uncertain objects in terms of their effect on the learning model. The useful objects can be used to improve the performance of the learning model. Conversely, the useless objects are those that have no positive effect on the learning model, and even make it worse. The unlabeled objects that cannot be determined to be useful or useless belong to uncertain. Intuitively, we can categorize each unlabeled object by the probability predicted by the learning model. However, in some cases, objects with different decisions could result in different risks. Therefore, we should take into consideration both the prediction probability and the decision risk to determine each unlabeled object.

In three-way decision, an object is determined to be positive, negative, or uncertain by using the idea of decision making with Bayesian minimum risk. A natural idea is to use the theory of three-way decision to evaluate unlabeled objects. But traditional three-way decision is a single view model. By integrating three-way decision with co-training, we propose a multi-view co-decision model to categorize unlabeled objects. For each unlabeled object, the co-decision results can be expressed as Table 3.

Table 3: Co-decision results by two views.

| | $a_P^2$ | $a_B^2$ | $a_N^2$ |
|---|---|---|---|
| $a_P^1$ | P | P | N |
| $a_B^1$ | P | B | P |
| $a_N^1$ | N | P | P |

In Table 3, $a_t^k$ denotes view k makes the decision t for an object x, where $k \in \{1,2\}$, $t \in \{P,B,N\}$, and P, B, and N in each cell denote the model with two views makes a co-decision to decide the object x to be positive, boundary, or negative, respectively.

In the proposed co-decision model, an acceptance decision is made when one of the two views confidently classifies the object as positive or negative and the other view as boundary; a rejection decision is made when one of the two views confidently determines the object to be positive and the other view to be negative; a wait-and-see decision can be only made when both views consider the object to be boundary. For the acceptance decision, since one of the two views is confident in its decision, the uncertain one could leverage the useful objects to improve its performance. For the rejection decision, two views are both confident in the decision, but their predictions are contradictory. The performance may deteriorate after learning from the divergent objects so that the co-decision model should discard this kind of unlabeled objects. For the wait-and-see decision, both views are unconfident to make a certain decision so that the co-decision model cannot use the uncertain unlabeled objects but keep them for further learning. For the unlabeled objects that both views are confident to determine to be positive or negative, the co-decision model can make an acceptance decision. However, considering that each view already has the ability to discern these objects, we do not consider them in order to simplify the learning process.

Taking into consideration both the decision and the risk, the collaborative decision costs under different actions can be described as:

$$
\begin{aligned}
&R(b_P|x) = min_{i \in \{P,N\} \wedge j \in \{B\}}\{R(a_i^1|x) + R(a_j^2|x), R(a_j^1|x) + R(a_i^2|x)\},\\
&R(b_B|x) = min_{i \in \{B\} \wedge j \in \{B\}}\{R(a_i^1|x) + R(a_j^2|x)\},\\
&R(b_N|x) = min_{i,j \in \{P,N\} \wedge i \neq j}\{R(a_i^1|x) + R(a_j^2|x)\},
\end{aligned}
\tag{11}
$$

where $R(b_P|x)$, $R(b_B|x)$, and $R(b_N|x)$ denote the costs for deciding an unlabeled object x to be useful, uncertain, or useless, respectively. According to Bayesian minimum risk decision, we can drive the following decision rules:

(P) if $R(b_P|x) < min\{R(b_B|x), R(b_N|x)\}$, then decide $x$ to be useful;
(B) if $R(b_B|x) < min\{R(b_P|x), R(b_N|x)\}$, then decide $x$ to be uncertain;
(N) if $R(b_N|x) < min\{R(b_P|x), R(b_B|x)\}$, then decide $x$ to be useless.

With the principle of the three-way co-decision, we can examine each unlabeled object and select some useful ones to improve the learning model. The process of the three-way co-decision model for partially labeled data can be depicted by Algorithm 3.

Algorithm 3 uses Algorithm 2 to decompose all condition attributes into two distinct reducts, on each of which a base classifier is trained on the initial labeled data. After initializing all parameters, the two classifiers repeatedly learn from each other by utilizing the useful objects determined by the three-way co-decision. More specifically, in each round of co-training, the performance of the two classifiers is evaluated on the initial labeled data, and then all unlabeled objects are grouped into three disjoint sets using the principle of three-way decision under multi-view, i.e., the useful, uncertain, and useless sets. When the performance of one classifier does not decrease, the classifier is retrained on a certain number of useful objects determined by the constrained inequality; otherwise, the classifier does not change. The algorithm terminates if neither classifier is updated, and the final classifier is generated by combining the two learned classifiers.

Assume that a partially labeled data consists of $|L|$ labeled and $|N|$ unlabeled objects described by $|C|$ attributes ($|U| = |L| + |N|$). The time cost of training a base classifier is almost $O(|C||U|)$. In each round of co-training, the two classifiers learn from each other on some useful objects. In the worst case, Algorithm 3 terminates after $|N|$ rounds of co-training. Thus, based on two distinct reducts of a given partially labeled data, the time cost of Algorithm 3 is at most $O(|C||U|^2)$, and its space cost is almost $O(|C||U|)$.

---

Algorithm 3 Three-way co-decision model for partially labeled data

Input:

A partially labeled data $PS=(U=L\cup N, A=C\cup D, V', f)$ and a confidence threshold parameter $\delta$;

Output:

A combined classifier H;

1: Decompose the condition attribute set C into two distinct semi-supervised reducts $P_1$ and $P_2$ by Algorithm 2;
2: Train two base classifiers $H_1$ and $H_2$ on L using the reducts $P_1$ and $P_2$, respectively;
3: Set the initial error rates and the sets of initial unlabeled objects for the two classifiers, $t\leftarrow 0$, $Err_1^t\leftarrow 0.5$, $Err_2^t\leftarrow 0.5$, $N_{P,1}^t\leftarrow\emptyset$, $N_{P,2}^t\leftarrow\emptyset$, $|N_{P,1}^t|\leftarrow 1$, $|N_{P,2}^t|\leftarrow 1$, $N^t=N$, $Update^t\leftarrow True$;
4: While $Update^t=True$ Do
5: Test the error rates $Err_1^{t+1}$ and $Err_2^{t+1}$ of the two classifiers $H_1$ and $H_2$ on L, $Update^{t+1}\leftarrow False$;
6: Categorize unlabeled data $N^t$ into the sets of useful objects $N_P^{t+1}$, uncertain objects $N_B^{t+1}$, and useless objects $N_N^{t+1}$ with the principle of the three-way co-decision;
7: Label each useful object with the class that one of the two classifiers confidently predicts, and update the unlabeled data $N^{t+1}\leftarrow N^t-N_N^{t+1}$;
8: If $Err_1^{t+1}<Err_1^t$ Then
9: Select the uncertain objects $N_{P,1}^{t+1}$ of $H_1$ from $N_P^{t+1}$;
10: Randomly pick a certain number of unlabeled objects $N_{P^*,1}^{t+1}$ from $N_{P,1}^{t+1}$ to keep the inequality $Err_1^{t+1}*|N_{P,1}^t\cup N_{P^*,1}^{t+1}|<Err_1^t*|N_{P,1}^t|$;
11: Retrain $H_1$ on $L\cup N_{P^*,1}^{t+1}$, $N_{P,1}^{t+1}\leftarrow N_{P,1}^t\cup N_{P^*,1}^{t+1}$, $Update^{t+1}\leftarrow True$;
12: End If
13: If $Err_2^{t+1}<Err_2^t$ Then
14: Select the uncertain objects $N_{P,2}^{t+1}$ of $H_2$ from $N_P^{t+1}$;
15: Randomly pick a certain number of unlabeled objects $N_{P^*,2}^{t+1}$ from $N_{P,2}^{t+1}$ to keep the inequality $Err_2^{t+1}*|N_{P,2}^t\cup N_{P^*,2}^{t+1}|<Err_2^t*|N_{P,2}^t|$;
16: Retrain $H_2$ on $L\cup N_{P^*,2}^{t+1}$, $N_{P,2}^{t+1}\leftarrow N_{P,2}^t\cup N_{P^*,2}^{t+1}$, $Update^{t+1}\leftarrow True$;
17: End If
18: $t\leftarrow t+1$;
19: End While
20: Combine the two classifiers into a final classifier $H=Combine(H_1,H_2)$;
21: Return the combined classifier $H$.

---

### 3.4. Theoretical analysis on the effectiveness of co-decision model

Considering the fact that the data in practical application typically has only a naturally undivided attribute set, the co-decision model relaxes the assumption of sufficient and redundant views in standard co-training into two distinct reducts. From the perspective of attribute reduction, each reduct is a jointly sufficient subset of all attributes that can preserve the overall discriminating power as the original attribute set. In addition, the algorithm for attribute reduction keeps two reducts to share common attributes as few as possible, and each reduct describes the data in different viewpoints such that the two trained classifiers in the co-decision model are sufficient and diverse to learn from each other. The researches in [20, 49] have shown that the process of co-training can succeed even if the two classifiers have a large diversity, which further guarantees that the proposed co-decision model could work well for partially labeled data.

The quality of unlabeled objects is another key factor for the success of co-training. On the one hand, the co-decision model employs the strategy of three-way decision to determine unlabeled objects to be useful, useless, or uncertain. In other words, the determination of each unlabeled object is not only related to the prediction probability, but also to the misclassification cost. On the other hand, after categorizing unlabeled data, some useful objects are selected for each classifier only when the estimated performance of the classifier does not deteriorate. Essentially, the principle of noise learning [1] is implicitly embedded into the co-decision model. In general, the performance of a classification model learned from noisy objects is constrained by the following equality:

$$m = \frac{c}{\epsilon^2(1-2\eta)^2}, \tag{12}$$

where $m$ is the number of objects for learning, $\epsilon$ is the worst-case error rate, $\eta(\eta < 0.5)$ is an upper bound on the classification noise rate, and $c$ is constant with respect to learning task.

By reforming the above equality, the following utility function with respect to the classification noise rate is obtained:

$$u = \frac{c}{(1-2\eta)^2} = m\epsilon^2. \tag{13}$$

To reduce the classification noise rate, the utility function should decrease in each iteration, i.e., $u^{t+1} < u^t$. The following inequality can be derived:

$$m^{t+1}(\epsilon^{t+1})^2 < m^t(\epsilon^t)^2. \tag{14}$$

Equivalently, we have

$$m^{t+1}\epsilon^{t+1} < m^t\epsilon^t, \tag{15}$$

and also the following constrained condition should be satisfied:

$$0 < \frac{\epsilon^{t+1}}{\epsilon^t} < \frac{m^t}{m^{t+1}} < 1. \tag{16}$$

According to (15) and (16), the inequality $m^{t+1}\epsilon^{t+1} < m^t\epsilon^t$ and the constraints $\epsilon^t < \epsilon^{t-1}$ and $m^{t-1} < m^t$ should be met simultaneously in each iteration.

In the proposed co-decision model, a classifier is considered for updating on some unlabeled data only when the estimated error rate does not increase. Furthermore, the classifier in each iteration only selects a certain number of unlabeled objects constrained by the inequality (15) in order to reduce (at least keep) the classification noise rate. Therefore, the co-decision model could use unlabeled data to improve its performance effectively.

Assume there are $n = |N|$ unlabeled objects in a given partially labeled data. The diversity of two classifiers on unlabeled data can be described by a confusion matrix (see Table 4).

Table 4: Diversity of two classifiers on unlabeled data.

| | $H_2$ positive | $H_2$ boundary | $H_2$ negative |
|---|---|---|---|
| $H_1$ positive | $n_{PP}$ | $n_{PB}$ | $n_{PN}$ |
| $H_1$ boundary | $n_{BP}$ | $n_{BB}$ | $n_{BN}$ |
| $H_1$ negative | $n_{NP}$ | $n_{NB}$ | $n_{NN}$ |

In the table, $n_{ij}$ denotes that the classifier 1 predicts an object to be i and the classifier 2 predicts the object to be j, where i and j belong to positive, boundary, or negative. In the first round of co-training, at most $n_{BP} + n_{BN}$ and $n_{PB} + n_{NB}$ unlabeled objects can be used to improve the classifier 1 and the classifier 2, respectively, so that total $n_{BP} + n_{BN} + n_{PB} + n_{NB}$ unlabeled objects could be utilized by the co-decision model. After each round of co-training, some uncertain unlabeled objects may become useful. As a result, the co-decision model could at most use $n_{BP} + n_{BN} + n_{PB} + n_{NB} + n_{BB}$ unlabeled objects to improve its performance.

## 4. Empirical analysis

The purpose of the experiments is twofold. One is to verify the effectiveness of the proposed attribute reduction algorithm for partially labeled data, i.e., Algorithm 1. The other is set out to show the performance of the proposed model compared to other semi-supervised learning models for partially labeled data. All experiments were carried out on a computer with Windows 10 operating system, Intel Xeon (R) CPU E5-2670 v3@2.30 GHz processor, and 32 GB Memory.

### 4.1. Investigated data sets and experiment design

Ten UCI data sets[1] are considered in the experiments, and the details are summarized in Table 5.

Table 5: Investigated data sets

| Data sets | $\|C\|$ | $\|U\|$ | $\|U/D\|$ | Missing | Inconsistency |
|---|---|---|---|---|---|
| credit-rating(credit) | 15(6) | 690 | 2 | Y | 8 |
| german-credit(german) | 20(7) | 1000 | 2 | N | 2 |
| gesture-phase-a2va3(gesture1) | 32(32) | 1260 | 5 | N | 27 |
| gesture-phase-b1va3(gesture2) | 32(32) | 1069 | 5 | N | 39 |

1. http://archive.ics.uci. edu/ml

| | | | | | |
|---|---|---|---|---|---|
| horse-colic(horse) | 22(7) | 368 | 2 | Y | 0 |
| kdd-synthetic-control(kdd) | 60(60) | 600 | 6 | N | 0 |
| parkinson-speech-train(parkinson) | 26(26) | 1040 | 2 | N | 57 |
| sonar(sonar) | 60(60) | 208 | 2 | N | 0 |
| tic-tac-toe(ttt) | 9(0) | 958 | 2 | N | 0 |
| wine(wine) | 13(13) | 178 | 3 | N | 0 |

In Table 5, the second column denotes the number of condition attributes, in which the number of numerical attributes is listed in the brackets. While the number of objects and classes in each data set is shown in the third and fourth columns, respectively. The fifth column indicates whether the data set has missing values or not, and the last column reports the number of inconsistent objects within the data set.

To facilitate the experiments, missing values in each data set are all completed by the mean (or mode) of the corresponding attribute. While the numerical attributes in each data set are discretized into categorical attributes since the proposed model is primarily developed for partially labeled data with categorical attributes. Due to the simplicity and popularity, the technique of equal frequency binning with three bins is employed to discretize numerical attributes into categorical ones [9]. In the experiments, 10-fold cross-validation is employed. More specifically, in each fold, 90% of objects are selected for the training set, and the remaining objects are used as the test set. For a given label rate, the training set is further randomly partitioned into a set of labeled objects $L$ and a set of unlabeled objects $N$. For instance, if there is a training set with 1000 objects, under the label rate $\theta = 10\%$, a labeled set of 100 objects and an unlabeled set of 900 objects will be generated in the experiments.

4.2. Attribute reduction for partially labeled data

To test the effectiveness of the proposed attribute reduction algorithm for partially labeled data, we conduct the experiments on all selected data sets under the label rate $\theta = 10\%$. In the proposed algorithm, a confidence parameter is needed to generate the discernibility matrix, which could provide the adaptability to noise. The higher the confidence threshold, the lower the degree of tolerance to noise. The confidence discernibility matrix degenerates into traditional discernibility matrix when the confidence threshold is set to 1. In practice, the setting for this parameter is task-specific and is suggested to select from the range (0.5, 1). For simplicity, we empirically set the confidence parameter $\delta$ to 0.75 in all experiments. The reduct information of all selected data sets is shown in Table 6.

Table 6: Results of semi-supervised attribute reduction under the label rate $\theta = 10\%$

| Data sets | Raw | Semi-supervised reduct | | | Ground-truth reduct | | | Approximate rate |
|---|---|---|---|---|---|---|---|---|
| | | Min | Max | Average | Min | Max | Average | |
| credit | 15 | 12 | 13 | 12.93 | 9 | 11 | 10.70 | 0.83 |
| german | 20 | 12 | 14 | 13.30 | 10 | 11 | 10.59 | 0.80 |
| gesture1 | 32 | 23 | 25 | 24.20 | 15 | 19 | 16.93 | 0.70 |
| gesture2 | 32 | 22 | 25 | 23.60 | 14 | 17 | 15.17 | 0.64 |
| horse | 22 | 11 | 15 | 13.67 | 8 | 10 | 8.99 | 0.66 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| kdd | 60 | 12 | 14 | 13.10 | 8 | 9 | 8.90 | 0.68 |
| parkinson | 26 | 19 | 20 | 19.80 | 15 | 16 | 15.65 | 0.79 |
| sonar | 60 | 8 | 9 | 8.23 | 6 | 7 | 6.57 | 0.80 |
| ttt | 9 | 8 | 8 | 8.00 | 8 | 8 | 8.00 | 1.00 |
| wine | 13 | 9 | 11 | 9.97 | 4 | 6 | 5.21 | 0.52 |
| Avg. | 28.9 | 13.6 | 15.4 | 14.68 | 9.70 | 11.4 | 10.67 | 0.73 |

In the table, we collect the reducts in 10-fold cross-validation. The statistical results, including the maximum, minimum, and average numbers of attributes in the reducts, are listed in the third to fifth columns. Besides, we also record the real reduct information for comparison, i.e., the reduct under the label rate $\theta = 100\%$. The difference between the semi-supervised reduct and the ground-truth supervised reduct is indicated in the last column, i.e., approximate rate, which is computed by the average number of attributes in the ground-truth reduct over that in the semi-supervised reduct.

In Table 6, it is evident that some of the attributes are removed from each data set after semi-supervised attribute reduction. By viewing the experimental results, we find that, in every fold of cross-validation, some attributes are excluded from the reducts, but at the same time some attributes are always included in the reducts. The main reason for this may be that these attributes are completely irrelevant or strongly relevant to classification task. Compared with the ground-truth reduct, the proposed algorithm achieves an approximate rate of 73% on all data sets. It is noteworthy that the semi-supervised reduct of data set "ttt" under the label rate $\theta = 10\%$ is exactly the same as the ground-truth supervised reduct obtained under the label rate $\theta = 100\%$. These results demonstrate the potential of the proposed attribute reduction algorithm for partially labeled data.

### 4.3. The effectiveness of the co-decision model

The proposed co-decision  is compared with classic semi-supervised methods, including self-training and co-training. Original self-training [20] is a self-taught algorithm with only one view. It trains a base classifier on initial labeled data and iteratively selects some confident unlabeled data with their predictions to retrain the base classifier until the stop condition is met. Co-training is a multi-view paradigm in disagreement-based methods, but its constraint on view is hard to satisfy because most of data sets do not have naturally partitioned views. Fortunately, the work in [20] showed that co-training can still benefit from unlabeled objects by randomly splitting the original attribute set into two subsets. Thus, in our experiments, we split the attributes in each data set into two disjoint sets with almost equal size. For fair comparison, self-training with two random split views is also investigated. Moreover, we record the initial performance of semi-supervised methods for comparison. The settings for all selected methods are shown in Table 7.

Table 7: Settings for all selected methods.

| Methods | View generation | Object selection |
|---|---|---|
| ST-1View | Original attribute set | Confidence level |

| | | |
|---|---|---|
| ST-2Views | Random split attribute subsets | Confidence level |
| CT-2Views | Random split attribute subsets | Confidence level |
| CD-2Views | Attribute reduction | Minimum risk |

In Table 7, ST-1View and ST-2Views denote the methods of self-training with one view and two views, respectively. While CT-2Views and CD-2Views stand for the standard co-training and the proposed co-decision method, respectively. To learn from partially labeled data, ST-1View, ST-2Views, and CT-2Views require the confidence threshold parameters to determine useful unlabeled objects. The proposed CD-2Views also needs to generate the confidence discernibility matrix based on a confidence threshold and categorize unlabeled objects by a pair of threshold parameters, while the latter is calculated from practical risk functions and task-specific. For simplicity and fair comparison, we use the same parameters ($\delta = 0.75$, $\alpha = 0.75$, $\beta = 0.55$) in all experiments. More specifically, ST-1View and ST-2Views will select the unlabeled objects whose confidence levels are greater than $\alpha$, and CT-2Views will use the unlabeled objects when the predicted confidence of one classifier is greater than $\alpha$ but the other classifier is less than $\beta$. While CD-2Views will use the confidence threshold $\delta$ to generate the discernibility matrix and the threshold parameters $\alpha$ and $\beta$ to determine the useful, uncertain, and unuseful unlabeled objects, respectively.

To investigate the effectiveness of the proposed method, two different base classifiers, namely J48 and Naive Bayes, are utilized in the experiments. When the label rate is set to $\theta = 10\%$, the results of the selected methods on all data sets are shown in Tables 8 and 9.

In Tables 8 and 9, the columns of "Initial" and "Final" denote the error rates of the selected method learned from initial labeled data and further improved by unlabeled data, respectively, and their results are averaged from 10-fold cross-validation. The column of "Improv." indicates the degree of improvement on performance, which can be computed by dividing the performance gain over the initial performance, and the column of "Max Performance" shows the error rates of the classifier trained on all training data with true labels, i.e., data set under the label rate $\theta = 100\%$. The best results among the selected methods are all boldfaced. The row of "Avg." in the table shows the average error rates of the selected methods across all data sets. Note that the performance of multi-view models is calculated by averaging all base classifiers.

From Tables 8 and 9, it is observed that, under the label rate $\theta = 10\%$, the performance of the selected algorithm is significantly different. Self-training with one view (ST-1View) achieves the best performance improvement on some data sets, such as "gesture2" (8.51%) in Table 8, "wine" (16.63%) in Table 9, but its performance become worse on most of other data sets. Self-training with two views (ST-2Views) benefits from the framework of multi-view and obtains relatively stable results, while the final performance still deteriorates after learning from unlabeled data on most data sets. Co-training with two views (ST-2Views) can learn from each other so that it could improve its performance by exploiting unlabeled data. However, it is also shown that, on some data sets, the performance of ST-2Views is almost unchanged or even become worse.

While co-decision with two views (CD-2Views), by carefully selecting useful unlabeled data, gains a performance improvement on most data sets. By averaging all results on the selected data sets, the final performance of CD-2Views using J48 and Naive Bayes is improved by 4.09% and 6.00%, respectively. Although the performance of ST-1View is also enhanced by 0.67% and 1.70%, respectively, its final performance is much worse than that of CD-2Views.

Table 8: Average performance of the selected methods using J48 classifier ($\theta = 10\%$).

| | ST-1View | | | ST-2Views | | | CT-2Views | | | CD-2Views | | | Max |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Initial | Final | Improv. | Initial | Final | Improv. | Initial | Final | Improv. | Initial | Final | Improv. | Performance |
| credit | 0.2086 | 0.1948 | 6.61% | 0.2326 | 0.2354 | -1.18% | 0.2326 | 0.2275 | 2.18% | 0.2074 | **0.1919** | **7.48%** | 0.1592 |
| german | 0.3335 | 0.3376 | -1.23% | 0.3413 | 0.3466 | -1.55% | 0.3413 | 0.3383 | **0.88%** | 0.3334 | **0.3319** | 0.45% | 0.3319 |
| gesture1 | 0.5341 | 0.5468 | -2.38% | 0.5302 | 0.5413 | -0.21% | 0.5302 | 0.5452 | -0.03% | 0.5294 | **0.5167** | **2.40%** | 0.4546 |
| gesture2 | 0.6043 | **0.5528** | **8.51%** | 0.5790 | 0.5922 | -2.27% | 0.5790 | 0.5875 | -1.46% | 0.5837 | 0.5552 | 4.88% | 0.3772 |
| horse | 0.2353 | 0.2358 | -0.22% | 0.2634 | 0.2615 | 0.72% | 0.2626 | 0.2626 | 0.00% | 0.2351 | **0.2329** | **0.91%** | 0.1948 |
| kdd | 0.3967 | 0.3983 | -0.42% | 0.3733 | 0.3750 | 0.45% | 0.3868 | 0.3850 | 0.46% | 0.3833 | **0.3367** | **12.17%** | 0.1440 |
| parkinson | 0.4606 | 0.4615 | -0.21% | 0.4621 | 0.4702 | -1.74% | 0.4615 | 0.4615 | 0.00% | 0.4567 | **0.4525** | **0.93%** | 0.4039 |
| sonar | 0.3773 | 0.3822 | -1.30% | 0.4053 | 0.3977 | **1.86%** | 0.3949 | 0.3949 | 0.00% | 0.3782 | **0.3729** | 1.40% | 0.2225 |
| ttt | 0.3178 | 0.3221 | -1.34% | 0.3425 | 0.3389 | 1.07% | 0.3425 | 0.3453 | -0.82% | 0.3199 | **0.3143** | **1.76%** | 0.1426 |
| wine | 0.3018 | 0.3058 | -1.33% | 0.2733 | **0.2728** | 0.19% | 0.2733 | 0.2819 | -3.13% | 0.2993 | 0.2737 | **8.54%** | 0.0975 |
| Avg. | 0.3770 | 0.3738 | 0.67% | 0.3803 | 0.3832 | -0.27% | 0.3805 | 0.3830 | -0.19% | 0.3726 | **0.3579** | **4.09%** | 0.2528 |

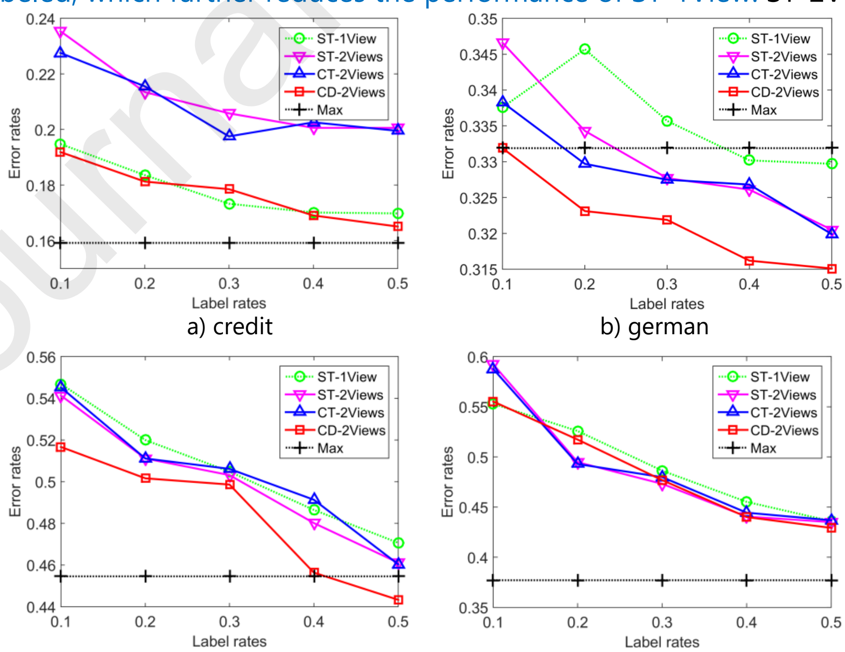
Table 9: Average performance of the selected methods using Navie Bayes classifier ($\theta = 10\%$).

| | ST-1View | | | ST-2Views | | | CT-2Views | | | CD-2Views | | | Max |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Initial | Final | Improv. | Initial | Final | Improv. | Initial | Final | Improv. | Initial | Final | Improv. | Performance |
| credit | 0.1464 | 0.1580 | -7.92% | 0.1464 | 0.1435 | 1.98% | 0.1526 | 0.1464 | **4.06%** | 0.1457 | **0.1407** | 3.40% | 0.1374 |
| german | 0.3137 | 0.3320 | -5.83% | 0.2890 | **0.2900** | **-0.35%** | 0.2890 | 0.2960 | -2.42% | 0.3000 | 0.3040 | -1.33% | 0.2554 |
| gesture1 | 0.4808 | 0.4857 | -1.02% | 0.4794 | 0.4786 | 0.17% | 0.4794 | 0.4762 | 0.66% | 0.4785 | **0.4508** | **5.80%** | 0.4229 |
| gesture2 | 0.6334 | 0.6268 | 1.04% | 0.6313 | 0.6315 | 0.00% | 0.6373 | 0.6319 | 0.84% | 0.6303 | **0.6118** | **2.94%** | 0.5605 |
| horse | 0.2474 | 0.2935 | -18.63% | 0.2396 | 0.2365 | 1.28% | 0.2366 | 0.2366 | 0.00% | 0.2355 | **0.2258** | **4.15%** | 0.2077 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| kdd | 0.1600 | 0.1473 | 7.94% | 0.1543 | 0.1517 | 1.73% | 0.1567 | 0.1507 | 3.83% | 0.1403 | **0.1050** | **25.18%** | 0.0707 |
| parkinson | 0.4673 | 0.4692 | -0.41% | 0.4702 | 0.4615 | **1.84%** | 0.4702 | 0.4712 | -0.20% | 0.4615 | **0.4558** | 1.25% | 0.3912 |
| sonar | 0.3467 | 0.4419 | -27.47% | 0.3474 | 0.3360 | **3.29%** | 0.3412 | 0.3412 | 0.00% | 0.3319 | **0.3267** | 1.58% | 0.2286 |
| ttt | 0.3704 | 0.3621 | -2.24% | 0.3710 | 0.3836 | -3.38% | 0.3810 | 0.3700 | 2.90% | 0.3721 | **0.3564** | **4.21%** | 0.2996 |
| wine | 0.1308 | 0.1091 | **16.63%** | 0.1356 | 0.1412 | -4.10% | 0.1456 | 0.1356 | 6.87% | 0.1124 | **0.0980** | 12.85% | 0.0515 |
| Avg. | 0.3297 | 0.3426 | 1.70% | 0.3264 | 0.3254 | 0.25% | 0.3289 | 0.3256 | 1.65% | 0.3208 | **0.3075** | **6.00%** | 0.2626 |

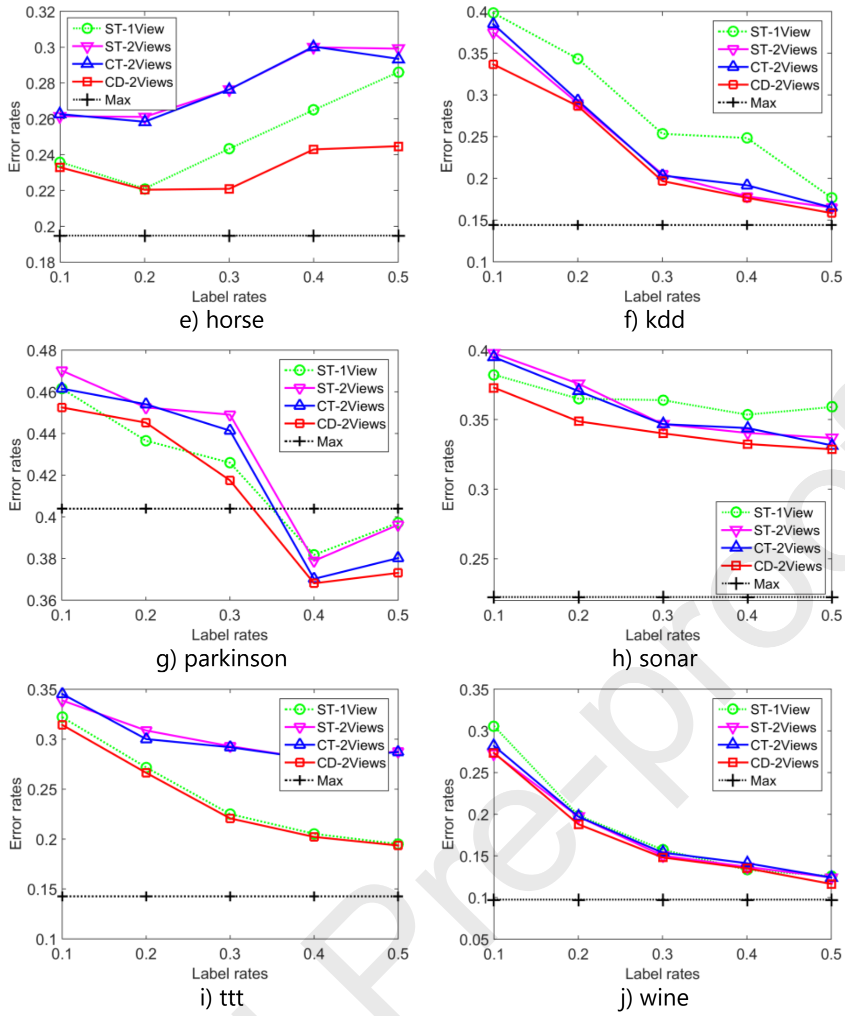
To fully evaluate the potential of the proposed model, some experiments under different label rates are also carried out. Their results are shown in Figures 2 and 3. Note that "Max" denotes the performance of the classifier under the label rate of 100%.

As shown in Figures 2 and 3, CD-2Views achieves impressive performance after capitalizing on unlabeled data. Since ST-1View is a single view model, unlabeled data can be only evaluated by itself. As a result, ST-1View obtains poor results on most data sets. For example, on data sets "german" and "horse", ST-1View under higher label rate even gets worse performance. One reason for these results may be the rarity of initial labeled data. Since the labeled objects are selected limitedly and randomly in the experiments, the generalization ability of the trained base classifier is relatively weak, resulting in the unstable performance, especially when the selected objects are not informative and representative. Moreover, the quality of unlabeled data used for learning has a considerable effect on performance. The self-labeled objects are inevitably mislabeled, which further reduces the performance of ST-1View. ST-2Views is



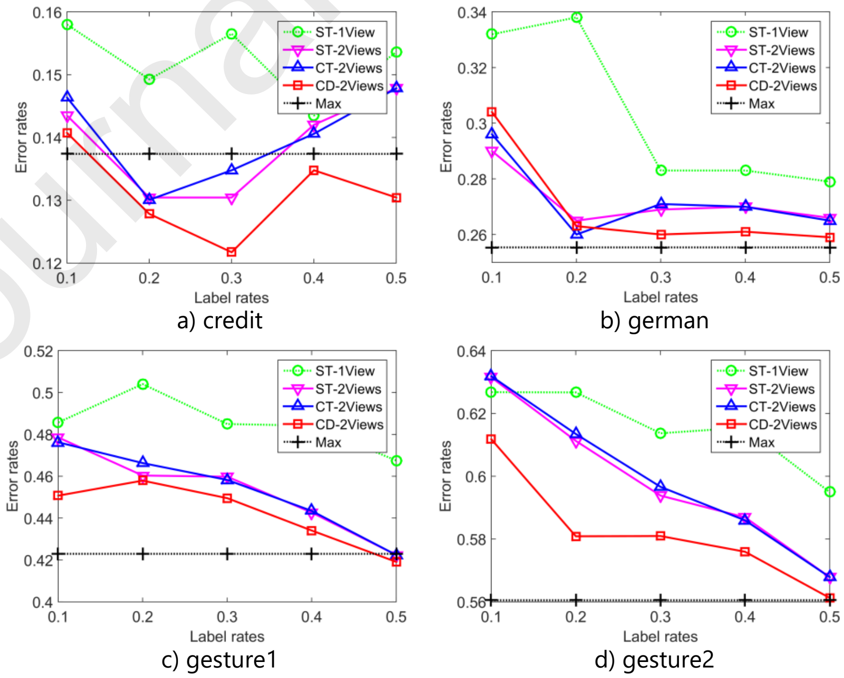

a) credit b) german

c) gesture1 d) gesture2




e) horse f) kdd g) parkinson h) sonar i) ttt j) wine

Fig 2: Average performance of the selected methods under different label rates (J48).




a) credit b) german c) gesture1 d) gesture2

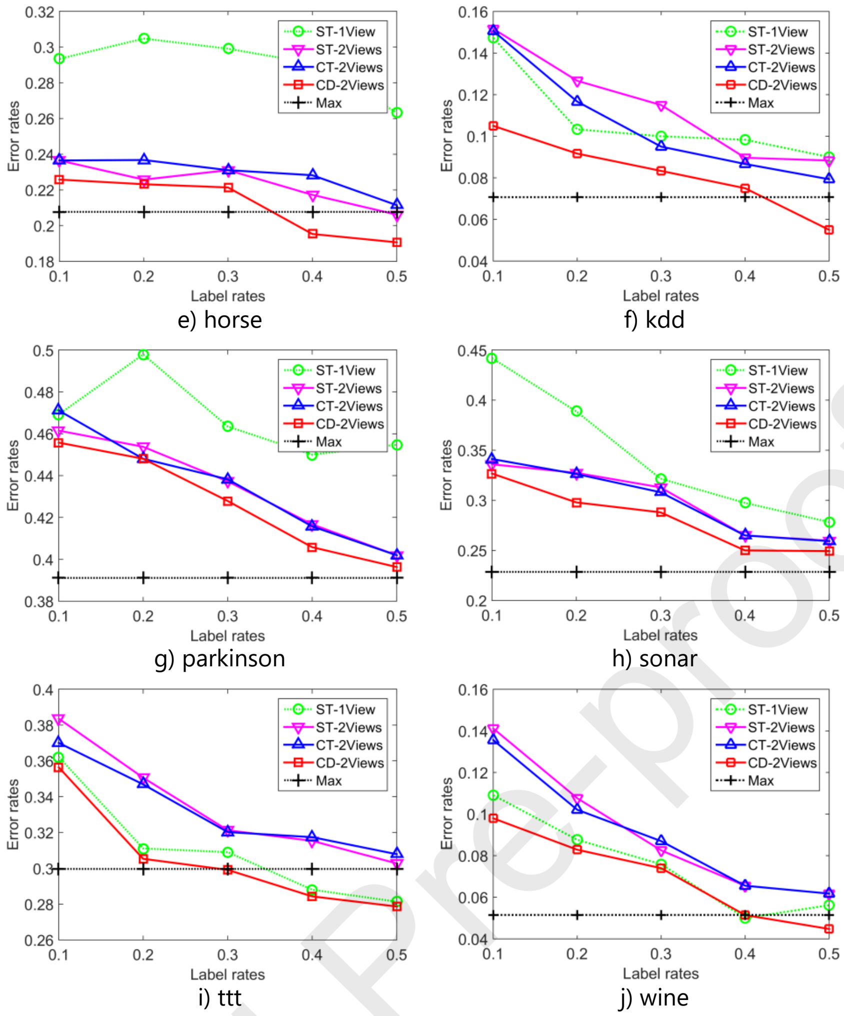



Fig 3: Average performance of the selected methods under different label rates (Naive Bayes).

a multi-view model. But its final performance is still unsatisfactory. In fact, the classifiers in ST-2Views are all self-taught. Furthermore, ST-2Views uses the randomly split attribute subsets to train its base classifiers. These reasons could attribute to the disappointing performance of ST-2Views. Although the classifiers in CT-2Views could use unlabeled data to improve the performance by learning from their counterparts, the subspaces for two classifiers are also randomly generated by halving the whole attribute set. Thus, the quality of the two classifiers cannot be guaranteed. As a result, some mislabeled objects may be selected by the two classifiers for their counterparts and the final performance of CT-2Views is undoubtedly poor. It can be verified by data sets "credit", "horse", and "ttt". Different from ST-2Views and CT-2Views, CD-2Views trains its base classifiers with reduct subspaces, each of which is a jointly sufficient subset of attributes that keeps the same level of discriminating power as the whole attribute set. Thus, the quality of the base classifiers in CD-2Views is much better than that of ST-2Views and CT-2Views. In addition, the performance of semi-supervised models is closely related to the unlabeled data used in the training stage. On the one hand, CD-2Views employs the theory of three-way decision to categorize unlabeled data in a collaborative way. Only the useful unlabled objects determined by the co-decision model are selected to learn, while the useless unlabeled objects will be directly abandoned by the model. On the other hand, the eligible unlabeled objects in each round of co-training are further tested by the effect on the performance of the classifier to learn. The training set of each classifier is updated only when the unlabeled objects to learn bring a positive effect on performance. With the above constraints, CD-2Views could use the really helpful unlabeled objects to improve the performance. On data sets "credit", "gesture2" and "parkinson", CD-2Views under some label rates achieves a slightly worse performance. These results may be due to the strict constraint on the number of useful unlabeled objects in each round of co-training so that the performance improvement is confined. However, on most of other data sets, CD-2Views under different label rates yields a significant performance improvement. These experimental results demonstrate that CD-2Views could effectively make use of unlabeled data to improve the performance, indicating the potential of the proposed model to learn from partially labeled data.

It is worth mentioning that, on some data sets, like "horse" and "parkinson" with J48, and "credit" and "german" with Naive Bayes, the performance of the selected methods decreases as the label rate increases. One possible explanation is that the scale of labeled data is not sufficient to train a classifier with good generalization ability. Besides, the methods cannot obtain satisfactory results when the initial labeled data is not representative. It is also impressive that the selected methods, especially the proposed one, achieve even better results than the maximum performance of data set, i.e., a trained classifier with the label rate $\theta = 100\%$. These findings are understandable because these methods benefit from unlabeled data and multi-view. These results confirm the fact that unlabeled data are helpful for improving learning performance.

## 5. Conclusions

Most real-world applications come with few labeled data and a large amount of unlabeled data. While the way of selecting and using informative unlabeled objects is of great importance to learning model for partially labeled data. In this paper, we develop the concept of the confidence discernibility matrix, based on which two semi-supervised attribute reduction algorithms are presented. To effectively learn from partially labeled data, we also introduce the co-decision model by incorporating the theory of three-way decision into co-training. Furthermore, the principle of noise learning is employed to conduct the selection of useful unlabeled data. The experimental results on UCI data sets show that the performance of our proposed model is promising when compared with the representatives. It should be noted that the proposed model focuses on partially labeled data with only categorical attributes so that the numerical attributes must be discretized. An extended model for partially labeled data with both categorical and numerical attributes is expected in the future. Also, the uncertainty analysis of the proposed model is also our future work.

## 6. Acknowledgement

The authors would like to thank Editor-in-Chief, editor, and anonymous reviewers for their valuable comments and helpful suggestions. The work was supported in part by the National Natural Science Foundation of China (Nos.61806127, 61703283, 61976145), in part by the Natural Science Foundation of Guangdong Province, China (Nos.2018A030310451, 2018A030310450), in part by Shenzhen Institute of Artificial Intelligence and Robotics for Society, and in part by the Bureau of Education of Foshan (Nos.2019XJZZ05).

## References

[ ] D. Angluin, P. Laird, Learning from noisy examples, Mach. Learn. 2 (1988) 343-370.
[2] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Proceedings of the 11th Annual Conference on Computational Learning Theory, Madison, Wisconsin, USA, 1998, pp. 92-100.
[3] D. Dai, H.X. Li, X.Y. Jia, X.Z. Zhou, B. Huang, S.N. Liang, A co-training approach for sequential three-way decisions, Int. J. Mach. Learn. Cybern. 11 (2020) 1129-1139.
[4] J.H. Dai, J.L. Chen, Feature selection via normative fuzzy information weight with application into tumor classification, Appl. Soft Comput. 92 (2020) 106299.
[5] J.H. Dai, H.F. Han, H. Hu, Q.H. Hu, J.H. Zhang, W.T. Wang, DualPOS: A semi-supervised attribute selection approach for symbolic data based on rough set theory, in: B. Cui, N. Zhang, J. Xu, X. Lian, D. Liu (Eds.) Proceedings of the 17th International Conference on Web-Age Information Management, Nanchang, China, 2016, pp. 392-402.
[6] J.H. Dai, H. Hu, W.Z. Wu, Y.H. Qian, D.B. Huang, Maximal-discernibility-pair-based approach to attribute reduction in fuzzy rough sets, IEEE Trans. Fuzzy Syst. 26 (2018) 2174-2187.
[7] J.H. Dai, Q.H. Hu, J.H. Zhang, H. Hu, N.G. Zheng, Attribute selection for partially labeled categorical data by rough set approach, IEEE Trans. Cybern. 47 (2017) 2460-2471.
[8] J.H. Dai, Q.H. Hu, H. Hu, D.B. Huang, Neighbor inconsistent pair selection for attribute reduction by rough set approach, IEEE Trans. Fuzzy Syst. 26 (2018) 937-950.
[9] E. Frank, M.A. Hall, I.H. Witten, The WEKA workbench, online appendix for "Data mining: Practical machine learning tools and techniques", Fourth Edition ed., Morgan Kaufmann, 2016.
[10] 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.
[ 1] R. Jensen, S. Vluymans, N. Mac Parthalain, C. Cornelis, Y. Saeys, Semi-supervised fuzzy-rough feature selection, in: Proceedings of the 15th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Tianjin, China, 2015, pp. 185-195.
[ 2] C.C. Kuo, H.L. Shieh, A semi-supervised learning algorithm for data classification, Int. J. Pattern Recognit. Artif. Intell. 29 (2015) 1551007.
[ 3] B.Y. Li, J.M. Xiao, X.H. Wang, Feature selection for partially labeled data based on neighborhood granulation measures, IEEE Access. 7 (2019) 37238-37250.
[ 4] P. Lingras, M. Chen, D.Q. Miao, Semi-supervised rough cost/benefit decisions, Fundam. Inform. 94 (2009) 233-244.
[ 5] C.H. Liu, K. C. Cai, D.Q. Miao, J. Qian, Novel matrix-based approaches to computing minimal and maximal descriptions in covering-based rough sets, Inf. Sci. 539 (2020) 312-326
[ 6] K.Y. Liu, E.C.C. Tsang, J.J. Song, H.L. Yu, X.J. Chen, X.B. Yang, Neighborhood attribute reduction approach to partially labeled data, Granul. Comput. 5 (2020) 239-250.
[ 7] K.Y. Liu, X.B. Yang, H.L. Yu, J.S. Mi, P.X. Wang, X.J. Chen, Rough set based semi-supervised feature selection via ensemble selector, Knowl.-Based Syst. 165 (2019) 282-296.

[18] D.Q. Miao, C. Gao, N. Zhang, Z.F. Zhang, Diverse reduct subspaces based co-training for partially labeled data, Int. J. Approx. Reason. 52 (2011) 1103-1117.
[19] F. Min, F.L. Liu, L.Y. Wen, Z.H. Zhang, Tri-partition cost-sensitive active learning through kNN, Soft Comput. 23 (2019) 1557-1572.
[20] K. Nigam, R. Ghani, Analyzing the effectiveness and applicability of co-training, in: Proceedings of the 9th ACM International Conference on Information and Knowledge Management, McLean, VA, 2000, pp. 86-93.
[21] N.M. Parthalain, R. Jensen, Fuzzy-rough set based semi-supervised learning, in: 2011 IEEE International Conference on Fuzzy Systems, Taipei, Taiwan, 2011, pp. 2465-2472.
[22] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (1982) 341-356.
[23] Z. Pawlak, Rough sets: Theoretical aspects of reasoning about data, Kluwer Academic Publishers, Dordrecht, Netherlands, 1991.
[24] 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.
[25] 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.
[26] C. Sengoz, S. Ramanna, Learning relational facts from the web: A tolerance rough set approach, Pattern Recognit. Lett. 67 (2015) 130-137.
[27] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, in: R. Slowinski (Ed.) Intelligent decision support: Handbook of applications and advances of the rough sets theory, Springer Netherlands, Dordrecht, 1992, pp. 331-362.
[28] I. Triguero, S. García, F. Herrera, Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study, Knowl. Inf. Syst. 42 (2015) 245-284.
[29] R. Wang, D.G. Chen, S. Kwong, Fuzzy-rough-set-based active learning, IEEE Trans. Fuzzy Syst. 22 (2014) 1699-1704.
[30] B. Yang, J.H. Li, Complex network analysis of three-way decision researches, Int. J. Mach. Learn. Cyb. 11 (2020) 973-987.
[31] X. Yang, T.R. Li, D. Liu, H.M. Chen, C. Luo, A unified framework of dynamic three-way probabilistic rough sets, Inf. Sci. 420 (2017) 126-147.
[32] Y.Y. Yao, Probabilistic rough set approximations, Int. J. Approx. Reason. 49 (2008) 255-271.
[33] Y.Y. Yao, Three-way decisions with probabilistic rough sets, Inf. Sci. 180 (2010) 341-353.
[34] Y.Y. Yao, The superiority of three-way decisions in probabilistic rough set models, Inf. Sci. 181 (2011) 1080-1096.
[35] Y.Y. Yao, Three-way decisions and cognitive computing, Cogn. Comput. 8 (2016) 543-554.
[36] Y.Y. Yao, Three-way decision and granular computing, Int. J. Approx. Reason. 103 (2018) 107-123.
[37] Y.Y. Yao, Three-way conflict analysis: Reformulations and extensions of the Pawlak model, Knowl.-Based Syst. 180 (2019) 26-37.
[38] Y.Y. Yao, Tri-level thinking: models of three-way decision, Int. J. Mach. Learn. Cybern. 11 (2020) 947-959.
[39] Y.Y. Yao, Three-way granular computing, rough sets, and formal concept analysis, Int. J. Approx. Reason. 116 (2020) 106-125.
[40] Y.Y. Yao, S. Wang, X.F. Deng, Constructing shadowed sets and three-way approximations of fuzzy sets, Inf. Sci. 412-413 (2017) 132-153.
[41] Y.Y. Yao, S.K.M. Wong, A decision theoretic framework for approximating concepts, Int. J. Man-Mach. Stud. 37 (1992) 793-809.
[42] Y.Y. Yao, B.X. Yao, Covering based rough set approximations, Inf. Sci. 200 (2012) 91-107.
[43] Y.Y. Yao, Y. Zhao, Attribute reduction in decision-theoretic rough set models, Inf. Sci. 178 (2008) 3356-3373.
[44] Y.Y. Yao, Y. Zhao, Discernibility matrix simplification for constructing attribute reducts, Inf. Sci. 179 (2009) 867-882.
[45] X.D. Yue, Y.F. Chen, D.Q. Miao, H. Fujita, Fuzzy neighborhood covering for three-way classification, Inf. Sci. 507 (2020) 795-808.

[46] Q.H. Zhang, Q. Xie, G.Y. Wang, A survey on rough set theory and its applications, CAAI Trans. Intell. Technol. 1 (2016) 323-333.
[47] J. Zhou, W. Pedrycz, C, Gao, Z.H. Lai, X.D. Yue, Principles for constructing three-way approximations of fuzzy sets: A comparative evaluation based on unsupervised learning, Fuzzy Sets Syst. 2020, doi: https://doi.org/10.1016/j.fss.2020.06.019.
[48] Z.H. Zhou, A brief introduction to weakly supervised learning, Natl. Sci. Rev. 5 (2018) 44-53.
[49] Z.H. Zhou, M. Li, Semi-supervised learning by disagreement, Knowl. Inf. Syst. 24 (2010) 415-439.
[50] X.J. Zhu, and A.B. Goldberg, Introduction to semi-supervised learning, Morgan & Claypool Publishers, Cambridge, MA, USA, 2009.

**Declaration of interests**

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Can Gao: Conceptualization, Methodology, Software, Data curation, Writing-Original draft preparation.

Jie Zhou: Methodology.

Duoqian Miao: Methodology, Reviewing and Editing.

Jiajun Wen: Software, Data analysis.

Xiaodong Yue: Software, Data analysis.