

Article

# Three-Way Co-Training with Pseudo Labels for Semi-Supervised Learning

Liuxin Wang<sup>1,2</sup>, Can Gao<sup>1,2,3,\*</sup> , Jie Zhou<sup>1,2,3</sup>  and Jiajun Wen<sup>1,2,3</sup>

<sup>1</sup> College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China; wx\_020306@163.com (L.W.); jie\_jpu@163.com (J.Z.); enjoy\_world@163.com (J.W.)

<sup>2</sup> Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen 518060, China

<sup>3</sup> SZU Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518060, China

\* Correspondence: 2005gaocan@163.com

**Abstract:** The theory of three-way decision has been widely utilized across various disciplines and fields as an efficient method for both knowledge reasoning and decision making. However, the application of the three-way decision theory to partially labeled data has received relatively less attention. In this study, we propose a semi-supervised co-training model based on the three-way decision and pseudo labels. We first present a simple yet effective method for producing two views by assigning pseudo labels to unlabeled data, based on which a heuristic attribute reduction algorithm is developed. The three-way decision is then combined with the concept of entropy to form co-decision rules for classifying unlabeled data into useful, uncertain, or useless samples. Finally, some useful samples are iteratively selected to improve the performance of the co-decision model. The experimental results on UCI datasets demonstrate that the proposed model outperforms other semi-supervised models, exhibiting its potential for partially labeled data.

**Keywords:** three-way decision; co-training; pseudo labels; normalized entropy; partially labeled data

**MSC:** 68T30



**Citation:** Wang, L.; Gao, C.; Zhou, J.; Wen, J. Three-Way Co-Training with Pseudo Labels for Semi-Supervised Learning. *Mathematics* **2023**, *11*, 3348. <https://doi.org/10.3390/math11153348>

Academic Editors: Weihua Xu, Jinhai Li and Xibei Yang

Received: 20 June 2023

Revised: 22 July 2023

Accepted: 23 July 2023

Published: 31 July 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In the era of information and data, there is an increasing need to process data described by a large number of attributes, such as genetic analysis, medical image classification, and text mining, which poses a huge challenge to traditional machine learning algorithms. Attribute reduction [1,2] has played an important role in removing irrelevant or redundant attributes, while retaining the most informative ones, which can improve computational efficiency and performance, and mitigate the risk of overfitting and the curse of dimensionality.

Rough set theory [3] has been proven to be a powerful mathematical tool for handling incomplete, uncertain, or imprecise data. Since Pawlak's groundbreaking research [4], the theory has undergone extensive development and application [5–7]. Presently, numerous attribute reduction methods based on rough set theory have been proposed, including discernibility matrix methods [3], positive region, and information entropy. Information entropy is one of the basic concepts of information theory. Pawlak et al. [8] introduced conditional entropy to assess the significance of attributes. Sun et al. [9] devised a novel neighborhood joint entropy from the algebraic and information perspectives of neighborhood rough sets. Jiang et al. [10] utilized the notion of relative decision entropy for the selection of informative attributes. Gao et al. [11,12] defined granularity-based maximum decision entropy to evaluate the significance of attributes. Yang et al. [13] proposed a pseudo-label neighborhood relation and re-defined both the neighborhood rough set and some corresponding measures. Yuan et al. [14] defined an uncertainty measure based on fuzzy complementary entropy and developed the attribute evaluation criteria of the maximal information, minimal redundancy, and maximal interactivity based on the proposed

uncertainty measure. Xu et al. [15] expanded the concept of information entropy to handle fuzzy incomplete systems.

As a generalization of rough set theory, the theory of three-way decision [16–25] is a decision-making methodology that introduces three options—acceptance, non-commitment, and rejection—by both considering the decision itself and the decision costs, rather than solely being a deterministic or binary decision. The three-way decision has now become one of the research hotspots in rough set theory, and there have been many related studies. Li et al. [26] proposed an axiomatic approach for describing three-way concepts using multiple granularities, and utilized the concept of set approximation to simulate cognitive processes. Wang et al. [27] proposed a solution to the problem of attribute reduction based on decision region preservation in rough sets. Ren et al. [28] introduced the three-way decision to the concept lattice and systematically studied the methods for the three-way concept lattice. Huang et al. [29] combined the distance function with the three-way decision to compute the neighborhood for mixed data. Zhang et al. [30] applied the three-way decision to the concept of the neighborhood information system and decomposed it into a three-level structure. Kong et al. [31] divided all attributes in the information table into three disjoint sets and developed a new granular structure for granular computing. Fang et al. [32] presented a framework that utilizes the three-way decision and discernibility matrix to address cost-sensitive approximation attribute reduction.

Typically, rough set-based approaches are employed for either fully labeled or unlabeled data. However, many real-world datasets may consist of both labeled and unlabeled data. To address the challenge of handling partially labeled data, several rough set-based methods have been developed. Miao et al. [33] defined a semi-supervised discernibility matrix and developed a novel rough co-training model to capitalize on unlabeled data to improve the performance of classifiers learned only from labeled data. By defining a novel discernibility matrix, Miao et al. [33] proposed a rough co-training model that harnesses unlabeled data to improve the accuracy of classifiers trained exclusively on labeled data. Wang et al. [34] applied Gaussian kernel-based similarity relation to evaluating the samples' inconsistency and developed an active learning model. To effectively deal with big data, Li et al. [35] used condition neighborhood granularity and neighborhood granularity to represent the importance of attributes, and provided an attribute reduction method for numerical attributes. Liu et al. [36] constructed multiple attribute fitness functions by the class local neighborhood decision error rate and used them to evaluate the importance of attributes. Xu et al. [6] developed a model and mechanism for a two-way learning system in fuzzy datasets based on information granules and developed an algorithm to implement different types of information granules. Pan et al. [37] defined a measure of semi-supervised neighborhood mutual information to generate the optimal semi-supervised reduct of partially labeled data by heuristic search. However, most of the aforementioned works focus on semi-supervised attribute reduction; less consideration is given to directly learning from semi-supervised models for partially labeled data. In fact, unlabeled data contain valuable information that can help the model better learn the information of data distribution and improve its generalization ability. However, noise and useless samples are also present in unlabeled data, which pose a significant threat to the learning process of partially labeled data. Therefore, it is important to develop a strategy that allows the model to efficiently select samples that are beneficial to itself. In this study, we propose a co-training model based on the three-way decision, and the main contributions are as follows:

- (1) To perform attribute reduction for partially labeled data, we propose a simple but effective labeling strategy for unlabeled data and develop a granular condition entropy-based heuristic attribute reduction algorithm for partially labeled data with pseudo labels.
- (2) To learn from unlabeled data effectively, we combine the three-way decision with information entropy into a co-training model and design three-way co-decision rules to classify unlabeled data into three sets of samples—useful, uncertain, and useless—which allows the model to learn from useful samples, thus improving performance.

- (3) To test the effectiveness of the proposed model, a large number of comparative experiments are conducted. The results demonstrate the superiority of the model and show its potential to handle partially labeled data.

The rest of this paper is ordered as follows. Section 2 briefly introduces the basic concepts of semi-supervised learning and three-way decision. Section 3 mainly describes the co-decision model based on the three-way decision. The experimental results and analysis are given in Section 4. Finally, Section 5 concludes the paper.

## 2. Preliminaries

This section briefly describes some related concepts in semi-supervised learning and three-way decision theory. More information can be found in [3,8,22,38–43].

### 2.1. Semi-Supervised Learning

In semi-supervised learning, partially labeled datum  $U$  containing  $l + n$  samples is divided into two parts: labeled data  $L = \{(x_i, y_i)\}_{i=1}^l$  and unlabeled data  $N = \{(x_j, ?)\}_{j=1}^{l+n}$  where  $U = L \cup N$  and  $l \ll n$ , so any sample in  $U$  is divided into  $L$  or  $N$  based on whether it has labels or not, that is, the intersection of  $L$  and  $N$  is  $\emptyset$ . Semi-supervised learning performs well in various machine learning tasks, such as semi-supervised clustering, semi-supervised classification, or semi-supervised regression. This paper focuses on the semi-supervised classification task [43].

Semi-supervised classification is a method that can use effective information in unlabeled data to improve the performance of supervised classifiers trained only on labeled data. It can generally be divided into four methods: generative methods, low-density separation methods, graph-based methods, and divergence-based methods [43]. Among them, co-training [40,41], as one of the more popular multi-view models in the divergence-based methods, has performed well on a large number of practical problems. It assumes that there are two views (attribute sets) to describe data, and two base classifiers trained separately on initially labeled data learn from each other using unlabeled data. Unfortunately, two independent and redundant views are difficult to guarantee in practical data. Therefore, it is essential to design a method to decompose the attribute set into two independent subsets to make the co-training model work well.

### 2.2. Three-Way Decision

In rough sets, an information system [3] represents the data that need to be processed, denoted as  $IS = (U, A, V, f)$ , where  $U$  is a non-empty set containing all samples;  $A$  is the set of attributes;  $V$  is the value domain of all attributes, so  $V_a$  represents the value domain of an attribute  $a$  in  $A$  and  $V = \cup V_a$ ; and  $f$  denotes a mapping function. For any sample  $x_i \in U$ , there exists a mapping relationship  $f(x_i, a) \in V_a$  for any attribute  $a \in A$ .

For any subset  $B$  of  $A$ ,  $U$  is divided into a set of equivalence classes  $U/B$ , and the equivalence class containing sample  $x$  is denoted as  $[x]_B$ . Let  $X$  be a subset of  $U$  and then the upper and lower approximations of  $X$  given  $B$  are defined as [3]

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

The  $\bar{B}(X)$  and  $\underline{B}(X)$  represent the set of samples that may belong to  $X$  and the set of samples that must belong to  $X$ , respectively. Particularly, the lower approximation  $\underline{B}(X)$  is also called the positive region  $POS_B(X)$  of  $X$  on  $U$ ; the difference between the upper and lower approximations of  $X$  is called the boundary region  $BND_B(X)$ , that is,  $BND_B(X) = \bar{B}(X) - \underline{B}(X)$ ; the set of samples outside the upper approximation of  $X$  is called the negative region, denoted as  $NEG_B(X) = U - \bar{B}(X)$ .

When the set of attributes  $A$  is further divided into the set of condition attributes  $C$  and the set of decision attributes  $D$ , the information system is called the decision information system, denoted as  $DS = (U, A = C \cup D, V, f)$ . Let  $U/D = \{Y_1, Y_2, \dots, Y_{|U/D|}\}$  be the

set of equivalence classes of the decision attribute set  $D$ , where  $Y_i$  is a set of samples with the decision  $i$ , then the positive region, the boundary region, and the negative region of  $D$  given  $C$  are defined as [3]

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

Let  $\Lambda = \{a_P, a_B, a_N\}$  be the set of behaviors that classify a sample into the positive region  $POS(X)$ , the boundary region  $BND(X)$ , or the negative region  $NEG(X)$ . Given a sample  $x \in U$ , the cost of taking different actions on  $x$  can be defined as [22]

$$\begin{aligned}
 R(a_P|[x]) &= \lambda_{PP}P(X|[x]) + \lambda_{PN}(1 - P(X|[x])), \\
 R(a_B|[x]) &= \lambda_{BP}P(X|[x]) + \lambda_{BN}(1 - P(X|[x])), \\
 R(a_N|[x]) &= \lambda_{NP}P(X|[x]) + \lambda_{NN}(1 - P(X|[x])),
 \end{aligned}
 \tag{3}$$

where  $P(X|[x])$  represents the probability that sample  $x$  belongs to  $X$ ,  $\lambda_{PP}$ ,  $\lambda_{BP}$ , and  $\lambda_{NP}$  are the costs of taking the action  $a_P$ ,  $a_B$ , or  $a_N$  when sample  $x$  is in  $X$ , respectively. Conversely, when the sample  $x$  is not in  $X$ , the costs of taking  $a_P$ ,  $a_B$ , or  $a_N$  are represented as  $\lambda_{PN}$ ,  $\lambda_{BN}$ , and  $\lambda_{NN}$ , respectively.

According to Bayesian minimal risk decision theory [39], the following rules can be obtained using the above decision costs [22]:

- (P) When  $R(a_P|[x]) \leq R(a_B|[x])$  and  $R(a_P|[x]) \leq R(a_N|[x])$ , classify sample  $x$  into the positive region;
- (B) When  $R(a_B|[x]) \leq R(a_P|[x])$  and  $R(a_B|[x]) \leq R(a_N|[x])$ , classify sample  $x$  into the boundary region;
- (N) When  $R(a_N|[x]) \leq R(a_P|[x])$  and  $R(a_N|[x]) \leq R(a_B|[x])$ , classify sample  $x$  into the negative region.

If we assume that the inequality  $(\lambda_{PN} - \lambda_{BN})(\lambda_{NP} - \lambda_{BP}) > (\lambda_{BP} - \lambda_{PP})(\lambda_{BN} - \lambda_{NN})$  holds, then the above rules can be further rewritten as [22]

- (P) When  $P(X|[x]) \geq \alpha$ , classify sample  $x$  into the positive region;
- (B) When  $\beta < P(X|[x]) < \alpha$ , classify sample  $x$  into the boundary region;
- (N) When  $P(X|[x]) \leq \beta$ , classify sample  $x$  into the negative region,

where  $\alpha = \frac{(\lambda_{PN} - \lambda_{BN})}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})}$ ,  $\beta = \frac{(\lambda_{BN} - \lambda_{NN})}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})}$ .

### 3. Three-Way Decision-Based Co-Training with Pseudo Labels

In this section, we first present the overall framework of the model proposed in this study. Then, we introduce a pseudo-labeling strategy to generate labels for partially labeled data and provide a heuristic attribute reduction algorithm. Finally, we propose a co-decision model to learn from unlabeled data.

#### 3.1. Overall Framework

Co-training [40,41] is a divergence-based multi-classifier model, in that two base classifiers are trained to learn from each other under two mutually independent views. However, there are no naturally divided two views in practical data, which limits the application of co-training. Moreover, base classifiers may learn from mislabeled or noisy samples, which worsens their performance. To address these problems, a three-way decision-based co-decision model is proposed in this study, and its overall framework is shown in Figure 1.

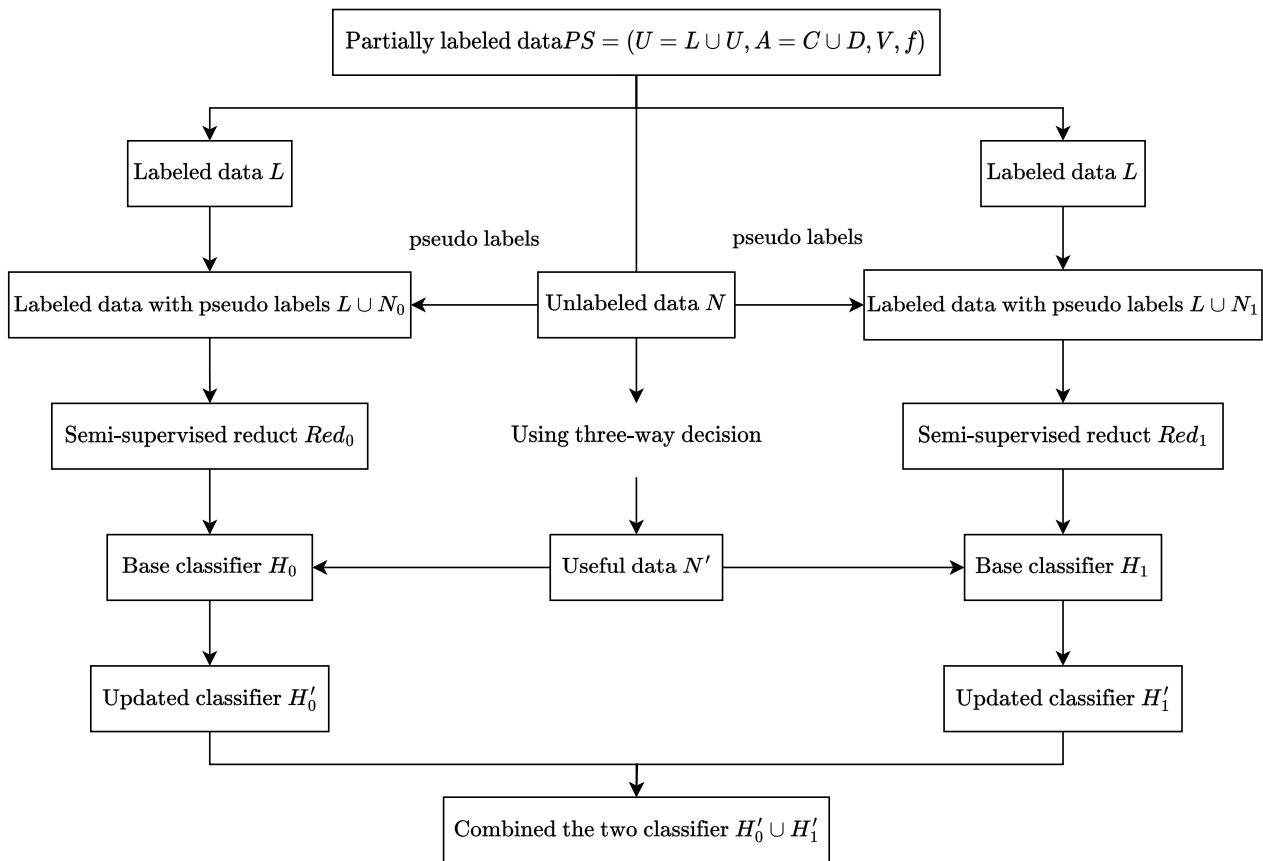


Figure 1. Framework of the three-way decision based co-training with pseudo labels.

Firstly, the unlabeled data are all tagged with the pseudo labels 0 or 1 and combined with the labeled data, respectively, which preserves the discriminative information of both the labeled and unlabeled data. Then, an attribute reduction algorithm is performed to obtain the optimal reduct on the two pseudo-labeled datasets, and two base classifiers are trained on the two reducts, respectively. Subsequently, the two classifiers are retrained iteratively on useful unlabeled data selected by using the three-way decision to make the classifiers only learn beneficial data and improve their performance until the stopping conditions are met. Finally, the two classifiers are combined to obtain the final classifier.

### 3.2. Semi-Supervised Attribute Reduction Based on Pseudo Labels

Traditional co-training has shown effectiveness in dealing with partially labeled data, but it remains an open question as to how to obtain two views from data with a naturally undivided attribute set. In this study, we propose a strategy that uses labeled samples and unlabeled samples with pseudo labels to form two views.

Assume a partially labeled datum  $PS = (U = L \cup N, A = C \cup D, V, f)$  has  $l$  samples in labeled datum  $L$  and  $n$  samples in unlabeled datum  $N$ , and  $|U| = u, u = l + n$ . Without loss of generality, assume that there are only two classes in the partially labeled data. For the unlabeled data  $N$ , we adopt a simple strategy of labeling all samples in  $N$  with pseudo labels 0 or 1, respectively, and generate two sets of pseudo-labeled data:  $N_0$  with pseudo label 0 and  $N_1$  with pseudo label 1. The two pseudo-labeled data are combined with the labeled datum  $L$  to form two views  $L'_0 = L \cup N_0$  and  $L'_1 = L \cup N_1$ . Factually, the generated pseudo-labeled data reflect the original partially labeled data from different perspectives, providing data diversity for the learning model. Formally, the partially labeled data with pseudo labels are represented as  $DS = (U = L \cup N'_k, A = C \cup D, V, f) (k = 0, 1)$ , where  $N'_k$  is the pseudo-labeled data after the labeling strategy.

In information theory [38], information entropy is represented by the expectation of information content from all possible events. Factually, it can be also used to quantify the uncertainty of attributes in a given dataset.

**Definition 1.** Given a partially labeled data  $PS = (U = L \cup N, A = C \cup D, V, f)$  and the partition  $U/B = \{X_1, X_2, \dots, X_{|U/B|}\}$  induced by the subset of condition attributes  $B \subseteq C$ , the information entropy of  $B$  over  $U$  is defined as [8]

$$H(B) = - \sum_{i=1}^{|U/B|} P(X_i) \log P(X_i), \tag{4}$$

where  $|\cdot|$  is the number of elements in a finite set and  $P(X_i) = \frac{|X_i|}{|U|}$ .

**Definition 2.** Given a partially labeled datum  $PS = (U = L \cup N, A = C \cup D, V, f)$  and the partitions  $U/B = \{X_1, X_2, \dots, X_{|U/B|}\}$ ,  $U/D = \{Y_1, Y_2, \dots, Y_{|U/D|}\}$  induced by the subset of condition attributes  $B \subseteq C$  and the decision attribute set  $D$ , respectively, the joint entropy between  $B$  and  $D$  is defined as [8]

$$H(B, D) = - \sum_{i=1}^{|U/B|} \sum_{j=1}^{|U/D|} P(X_i, Y_j) \log P(X_i, Y_j), \tag{5}$$

**Definition 3.** Given a partially labeled datum  $PS = (U = L \cup N, A = C \cup D, V, f)$  and the partitions  $U/B = \{X_1, X_2, \dots, X_{|U/B|}\}$ ,  $U/D = \{Y_1, Y_2, \dots, Y_{|U/D|}\}$  induced by the subset of condition attributes  $B \subseteq C$  and the decision attribute set  $D$ , respectively, the conditional entropy of  $D$  given  $B$  is defined [42]:

$$H(D|B) = - \sum_{i=1}^{|U/B|} \sum_{j=1}^{|U/D|} P(X_i, Y_j) \log P(Y_j|X_i), \tag{6}$$

where  $P(Y_j|X_i) = \frac{|X_i \cap Y_j|}{|X_i|}$ .

In rough sets, the universe under a subset of attributes can be divided into a set of information granules, each of which consists of some indiscernible samples. Factually, the size of the information granules reflects the discriminating power of the attribute subset. The finer the information granularity, the better the quality of the attribute subset.

**Definition 4.** Given a partially labeled datum  $PS = (U = L \cup N, A = C \cup D, V, f)$  and the partition  $U/B = \{X_1, X_2, \dots, X_{|U/B|}\}$  induced by the subset of condition attributes  $B \subseteq C$ , the granularity on  $U$  given  $B$  is defined as [44]

$$G(B) = - \sum_{i=1}^{|U/B|} P(X_i)^2. \tag{7}$$

**Definition 5.** Given a partially labeled datum  $PS = (U = L \cup N, A = C \cup D, V, f)$  and the partitions  $U/B = \{X_1, X_2, \dots, X_{|U/B|}\}$ ,  $U/D = \{Y_1, Y_2, \dots, Y_{|U/D|}\}$  induced by the subset of condition attributes  $B \subseteq C$  and the decision attribute set  $D$ , respectively, the granular conditional entropy of  $D$  given  $B$  is defined as [45]

$$GH(D|B) = - \sum_{i=1}^{|U/B|} P(X_i)^2 \sum_{j=1}^{|U/D|} P(X_i, Y_j) \log P(Y_j|X_i). \tag{8}$$

For any condition attribute subset  $B$ , its granular conditional entropy to  $D$  integrates the granularity and conditional entropy, which not only evaluates the quality of partitions induced by the condition attribute subset but also accumulates the uncertainty of each

condition class under different decisions, providing a better measure for the importance of an attribute subset.

**Definition 6.** Given a partially labeled datum  $PS = (U = L \cup N, A = C \cup D, V, f)$  and the partition  $U/B = \{X_1, X_2, \dots, X_{|U/B|}\}$  induced by the subset of condition attributes  $B \subseteq C$ , the importance of an attribute  $a \in (B - C)$  is defined as

$$sig(a, B, D) = H(D|B) - H(D|B \cup \{a\}). \tag{9}$$

Based on the pseudo-labeling strategy and granular conditional entropy, a semi-supervised attribute reduction algorithm is developed using the forward heuristic search. The procedure is described in Algorithm 1.

---

**Algorithm 1** Semi-supervised attribute reduction based on granular condition entropy.

---

**Input:** Partially labeled data  $PS = (U = L \cup N, A = C \cup D, V, f)$ .

**Output:** Semi-supervised reducts  $RED_0$  and  $RED_1$ .

- 1: Generate two pseudo-labeled data  $N_0$  and  $N_1$  from the unlabeled data using the pseudo labels 0 and 1, respectively;
  - 2: **for**  $k \in \{0, 1\}$  **do**
  - 3:    Compute the granular conditional entropy of  $GH(D|C)$  on the data  $L \cup N'_k$ ;
  - 4:    Calculate the granular conditional entropy  $GH(D|\{a_i\})(a_i \in C)$  to evaluate each attribute and add the attribute  $a_{opt}$  that has the lowest granular conditional entropy to  $RED_k$ ;
  - 5:    **while**  $GH(D|RED_k) \neq GH(D|C)$  **do**
  - 6:        Calculate the importance of each attribute  $sig(a_i, RED_k, D)$ ;
  - 7:        Select the attribute  $a_{opt}$  with highest importance;
  - 8:         $RED_k = RED_k \cup \{a_{opt}\}$ ;
  - 9:    **end while**
  - 10: **end for**
  - 11: **Return** Two semi-supervised reducts  $RED_0$  and  $RED_1$ .
- 

In Algorithm 1, two pseudo-labeled data are first generated. Then, on each, the overall granular conditional entropy under all condition attributes is calculated, and the attributes that have the lowest granular conditional entropies are iteratively selected to the reduct. The algorithm terminates when the granular conditional entropy of the obtained reduct is the same as that of all condition attributes. Finally, two optimal semi-supervised reducts are returned.

Assume that the partially labeled data have  $|U|$  samples and  $|C|$  condition attributes. The time cost of determining the optimal attribute is  $O(|C||U|^2)$  in each round of iteration. Note that the worst case is that it takes  $|C|$  rounds to select the optimal attribute in each iteration, then the worst time cost is  $O(|C|^2|U|^2)$ . The space cost is  $O(|C||U|)$  because it is necessary to store all partially labeled data.

### 3.3. Three-Way Co-Training Model for Partially Labeled Data

Co-training is a semi-supervised learning model that leverages unlabeled samples to enhance the performance of two classifiers. Therefore, the selection of unlabeled samples is crucial. Typically, each unlabeled sample can be classified as useful, uncertain, or useless. When the classifiers select useful unlabeled samples for learning, their classification performance will be improved; when the classifiers select useless samples, there may be negative effects on their performance; and uncertain samples indicate that the classifiers cannot classify them undoubtedly at this time, but they may become useful unlabeled samples in the next round of iterations. Therefore, our primary goal is to select useful samples and exclude useless ones as much as possible to improve the classifiers' performance. To accurately assess the confidence level of a classifier's prediction probability, we introduce the concept of normalized entropy.

**Definition 7.** Given a partially labeled datum  $PS = (U = L \cup N, A = C \cup D, V, f)$  and the probability distribution  $P^k(x) = \{p_1^k(x), p_2^k(x), \dots, p_{|U/D|}^k(x)\}, k \in \{0, 1\}$  predicted by the classifier for  $x \in N$  in the view  $k$ , the normalized entropy of the classifier on  $x$  is defined as

$$NE^k(x) = \frac{1}{\log|U/D|} \sum_{i=1}^{|U/D|} -p_i^k(x) \log p_i^k(x) \tag{10}$$

Normalized entropy uses the prediction probability distribution in a given view to reflect the degree of certainty of the classifier on its prediction. The higher the normalized entropy, the lower the credibility of the classifier in making the prediction on an unlabeled sample. Conversely, the lower the normalized entropy, the higher the credibility of the classifier.

The three-way decision is an effective method for making three optional decisions under uncertainty—acceptance, wait-to-see, and rejection—which coincides with our desire to classify unlabeled samples as useful, uncertain, and useless. However, the three-way decision is often used for a single classifier, so we propose a three-way co-decision model to classify unlabeled samples. In the model, each classifier divides a sample into useful, uncertain, or useless by using the Bayesian minimum risk theory. When the two classifiers both confidently classify an unlabeled sample as useful, uncertain, or useless, a co-decision  $P, B,$  or  $N$  is made. However, when one of the classifiers classifies the sample as uncertain and the other classifies the sample as useless, we consider the sample useless and make an  $N$  decision; when one of the classifiers classifies the sample as uncertain and the other classifies the sample as useful, we decide to make a  $P$  decision. However, when the predictions of two classifiers conflict, that is, one classifier considers the sample as useful and the other classifier classifies it as useless, we need to make further consideration to find as many useful samples as possible to improve the performance of classifiers, even though this may degrade the performance of one classifier but improve the performance of the combined classifier. Specifically, we use two threshold parameters  $\delta$  and  $\epsilon$  to evaluate the average normalized entropy of two classifiers. If the average normalized entropy is less than or equal to  $\delta$ , the sample is considered helpful to improve the performance of the combined classifier, so it is classified as  $P$ . If the average normalized entropy is greater than  $\delta$  and less than or equal to  $\epsilon$ , it means that the two classifiers generally do not have enough confidence in the sample and need further learning to judge it, so the sample is classified as  $B$ . If the average normalized entropy is greater than  $\epsilon$ , which indicates that the sample has only a negative impact on the performance of the combined classifier, it should be classified as  $N$ . Bearing this in mind, we further define the following three-way co-decision rules when two classifiers highly contradict each other:

- (P)** If  $\frac{1}{2}(NE^0(x) + NE^1(x)) \leq \delta$ , then determine sample  $x$  to  $P$ .
- (B)** If  $\delta < \frac{1}{2}(NE^0(x) + NE^1(x)) \leq \epsilon$ , then determine sample  $x$  to  $B$ .
- (N)** If  $\frac{1}{2}(NE^0(x) + NE^1(x)) > \epsilon$ , then determine sample  $x$  to  $N$ .

By using the normalized entropy and three-way decision, the co-decision results for each unlabeled sample can be expressed in Table 1.

**Table 1.** Co-decision rules by the proposed model.

	$a_P^1$	$a_B^1$	$a_N^1$
$a_P^0$	$P$	$P$	$Co - TWD$
$a_B^0$	$P$	$B$	$N$
$a_N^0$	$Co - TWD$	$N$	$N$

In Table 1,  $a_w^k$  indicates that the classifier  $k$  makes the decision  $w$ , where  $k \in \{0, 1\}$  and  $w \in \{P, B, N\}$ , the  $P, B,$  and  $N$  denote the co-decision result of two classifiers for an unlabeled sample, while the  $Co - TWD$  represents the co-decision result further determined by using the average normalized entropy and the three-way decision when two classifiers make different decisions.



Using these rules, unlabeled samples can be classified as useful, uncertain, and useless, from which only the useful samples are selected to retrain the classifiers. Algorithm 2 describes the procedure of the co-training model.

Algorithm 2 begins with Algorithm 1 to generate two semi-supervised reducts using pseudo labels. Two base classifiers are then trained separately on the two reducts. After initializing all parameters required for the co-training process, the unlabeled samples are iteratively classified into useful, uncertain, or useless by using co-decision rules. The classifiers can only be updated when useful samples exist. If the inequality constraint is satisfied, i.e., the classifier’s performance does not deteriorate after adding unlabeled samples, some useful samples are selected for updating the classifier in descending order of the average normalized entropy. The algorithm stops when neither classifier can be further updated, and the final classifier is obtained by combining the two classifiers.

---

**Algorithm 2** Co-training model for partially labeled data based on the three-way decision.

---

**Input:** A partially labeled data  $PS = (U = L \cup N, A = C \cup D, V, f)$ .

**Output:** A combined classifier  $H_{combined} = H_0^t \cup H_1^t$ .

- 1: Using Algorithm 1 to generate two pseudo-labeled data  $N'_0$  and  $N'_1$  and use them to obtain semi-supervised reducts  $RED_0$  and  $RED_1$ ;
  - 2: Train base classifiers  $H_0$  and  $H_1$  using  $RED_0$  and  $RED_1$ , respectively;
  - 3: Set the error rates, unlabeled samples, useful samples, and update flags for each classifier.  $Err_k^t = 0.5, N^t = N, N_{P,k}^t = \emptyset, Update_k^t = True, t = 0, k \in \{0, 1\}$ ;
  - 4: **while**  $Update_0^t = True$  or  $Update_1^t = True$  **do**
  - 5:      $Update_0^t = Update_1^t = False$ ;
  - 6:     Divide the unlabeled data  $N^t$  into useful samples  $N_P^{t+1}$ , uncertain samples  $N_B^{t+1}$ , and useless samples  $N_N^{t+1}$  by using the three-way decision;
  - 7:     **if**  $N_P^{t+1} \neq \emptyset$  **then**
  - 8:         Sort the samples in descending order in  $N_P^{t+1}$  based on the average normalized entropy of two classifiers  $H_0^t$  and  $H_1^t$ ;
  - 9:         **for**  $k \in \{0, 1\}$  **do**
  - 10:             Pick a certain number of samples  $N_{P^*,k}^{t+1}$  from  $N_P^{t+1}$  to ensure that the inequality  $Err_k^{t+1} * |N_{P,k}^t \cup N_{P^*,k}^{t+1}| < Err_k^t * |N_{P,k}^t|$  holds;
  - 11:              $N_{P,k}^t = N_P^{t+1} \cup N_{P^*,k}^{t+1}, Update_k^t = True$ ;
  - 12:             **end for**
  - 13:              $N^t = N^t - N_N^{t+1} - N_{P^*,0}^{t+1} \cup N_{P^*,1}^{t+1}$ ;
  - 14:         **end if**
  - 15:         **for**  $k \in \{0, 1\}$  **do**
  - 16:             **if**  $Update_k^t = True$  **then**
  - 17:                 Retrain classifier  $H_k^t$  on  $L \cup N_{P,k}^t$ ;
  - 18:             **end if**
  - 19:         **end for**
  - 20:          $t = t + 1$ ;
  - 21:     **end while**
  - 22: **Return** the combined classifiers  $H_{combined} = H_0^t \cup H_1^t$ .
- 

Assume a partially labeled datum has  $|U|$  samples and  $|C|$  condition attributes, where  $|L|$  samples are labeled and  $|N|$  samples are unlabeled. The time complexity of training each base classifier is  $O(|C||U|)$ . In the worst-case scenario, if there is only one useful unlabeled sample learned by one classifier in each iteration, it takes  $|N|$  iterations. Therefore, the time complexity of Algorithm 2 is  $O(|C||U|^2)$ , and the space complexity is  $O(|C||U|)$ .

#### 4. Empirical Analysis

In this section, we first test the effectiveness of semi-supervised attribute reduction and then the proposed model is compared with other semi-supervised learning methods. All

experiments are conducted on a computer with Windows 10 operating system configured with Inter(R) Core(TM) i7-7700K CPU @ 4.20 GHz 4.20 GHz, 16 GB RAM.

#### 4.1. Investigated Data Sets and Experiment Design

In the experiments, 16 UCI datasets are used in this experiment, and their details are presented in Table 2. The second column in Table 2 indicates the number of condition attributes, with the number of continuous attributes shown in brackets. The third and fourth columns display the sample size and number of classes for each dataset, respectively. The fifth column indicates whether the dataset has missing values.

**Table 2.** Investigated datasets.

Dataset Name	Number of Attributes	Number of Samples	Number of Classes	Missing Data
biodegradation (biode)	41 (41)	1055	2	N
cardiotocography (cardio)	21 (21)	2126	10	N
cmc (cmc)	9 (2)	1473	3	N
frogs (frogs)	22 (22)	7195	10	N
hepatitis (hepatitis)	19 (6)	155	2	Y
hungarian (hungarian)	13 (6)	294	2	Y
kr-vs-kp (krvskp)	36 (0)	3196	2	N
lymph (lymph)	18 (3)	148	4	N
newcylinder-bands(newcylinder)	37 (18)	540	2	Y
pima (pima)	8 (8)	768	2	N
quality-assessment-green (green)	62 (62)	98	2	N
spectf (spectf)	44 (44)	269	2	N
tic-tac-toe (ttt)	9 (0)	958	2	N
vehicle (vehicle)	18 (18)	846	4	N
vowel (vowel)	13 (10)	990	11	N
wine (wine)	13 (13)	178	3	N

In the experiments, missing values in each dataset are replaced by the mean or mode of their corresponding attributes. Continuous attribute values are first normalized to the range of [0, 1], and then an equal-frequency discretization technique with five bins is used [46]. To accurately evaluate the performance of the selected methods, we use a 10-fold cross-validation technique. For example, suppose there are 1000 samples in the partially labeled data, with the class distribution of 30% positive samples and 70% negative samples. In each fold of cross validation, 90% of the samples is randomly selected as the training set, while the remaining 10% is used as the test set, i.e., 900 training samples and 100 test samples are generated, and the original class distribution (30%, 70%) is retained. Considering a label rate of 10%, only 90 of the 900 training samples are selected as labeled samples, and the remaining 810 are treated as unlabeled samples. Finally, the average performance obtained from 10 cross validations is used as the final performance of the method for the given dataset.

#### 4.2. Attribute Reduction for Partially Labeled Data with Pseudo Labels

The semi-supervised attribute reduction based on the granular conditional entropy is used in the experiments. Specifically, the method uses a heuristic algorithm to generate the optimal reduct of partially labeled data by combining the information granularity with conditional entropy. The results of the attribute reduction at a label rate of 10% are shown in Table 3, where the second column shows the number of attributes of the dataset before attribute reduction, and the third to fifth columns are the maximum, minimum, and average number of remaining attributes after attribute reduction in 10 cross validations, respectively. The sixth column displays the number of attributes obtained after attribute reduction at a label rate of 100%, that is, all data are labeled.

**Table 3.** Results of semi-supervised attribute reduction based on granular conditional entropy.

Dataset Name	Raw	Reducts			Ground Truth
		Max	Min	Avg	
biode	41	15	12	13.8	12
cardio	21	8	5	6.3	4
cmc	9	9	8	8.3	7
frogs	22	15	14	14.4	10
hepatitis	19	11	9	10.1	9
hungarian	13	4	3	3.2	3
krvskp	36	32	30	31.2	29
lymph	18	8	7	7.5	6
newcylinder	37	17	16	16.7	15
pima	8	6	5	5.6	4
green	62	29	28	28.8	26
spectf	44	14	13	13.2	10
ttt	9	9	8	8.2	8
vehicle	18	14	12	13.2	10
vowel	13	11	10	10.9	10
wine	13	5	4	4.4	4
avg.	23.9	12.9	11.5	12.2	10.4

By observing Table 3, it can be found that after attribute reduction, the completely irrelevant and some redundant attributes are excluded from the obtained attribute subsets, thus reducing the redundant information as much as possible while preserving the inherent information of the dataset. At the same time, on the dataset “biode”, “hepatitis”, and “tic”, the minimum number of attributes in the reduct is nearly equivalent to that of the GT (ground truth), which implies that the proposed method can achieve the attribute reduction performance as the fully supervised method. The proposed method achieved an average attribute reduction rate of 48.95% across all selected datasets, demonstrating its potential in reducing the number of attributes required for classification.

#### 4.3. Effectiveness of the Proposed Co-Training Model

To demonstrate the performance of the proposed model, it is compared with classical semi-supervised methods, including self-training, co-training, and their extensions.

The classic self-training is a self-learning model. It first trains a basic classifier on labeled data, followed by iteratively selecting some confident samples from unlabeled samples to learn until the stop condition is met. Co-training is a multi-view model, in which two classifiers learn from each other on unlabeled data, but it has to fulfill the condition that two views must be sufficient and independent. Nevertheless, such a condition is usually difficult to satisfy in practical problems. Fortunately, the work of Nigam et al. [47] demonstrated that even if the raw data are randomly split into two attribute subsets, the co-training classifier can still learn from unlabeled samples. Therefore, we divided each condition attribute set of the dataset into two disjoint subsets by half-splitting attributes. In addition, for a more comprehensive comparison, we set the self-training into two cases: self-training with a single view and self-training with two randomly divided views. Moreover, we evaluated single-view self-training in two cases: data after attribute reduction and data without attribute reduction. The settings for these models are shown in Table 4.

**Table 4.** Settings of comparison methods.

Methods	Generated Views
ST-1V	Original attribute set
ST-1VR	Attribute reduction
ST-2V	Random split attribute subsets
CT-2V	Random split attribute subsets
CT-TWD	Attribute reduction with pseudo-labeled data

In Table 4, ST-1V and ST-2V denote the single-view self-training and two-view self-training, respectively, and ST-1VR denotes the single-view self-training after attribute reduction. In order to comprehensively compare the performance of the proposed model, we adopt a semi-supervised neighborhood discriminant index, which is a filter method that combines the supervised neighborhood discriminant index with unsupervised Laplacian information. CT-2V represents the classical co-training, while CT-TWD denotes the proposed three-way co-training model. To learn useful unlabeled samples, a threshold parameter needs to be set for ST-1V, ST-1VR, ST-2V, and CT-2V, and the model proposed in this study requires two pairs of parameters, where the first pair can be obtained based on the Bayesian minimum risk decision, while the second pair is calculated by the defined normalized entropy. For a simple and fair comparison, these parameters are all empirically set to  $\alpha = 0.75$ ,  $\beta = 0.55$ ,  $\delta = 0.80$ , and  $\varepsilon = 0.95$ . For ST-1V and ST-2V, the unlabeled sample with a prediction probability greater than  $\alpha$  is selected for learning. For CT-2V, the unlabeled sample is used for learning when its prediction probability of one classifier is greater than  $\alpha$ , and the probability predicted by the other classifier is less than  $\beta$ . For the CT-TWD in this study, the thresholds  $\alpha$  and  $\beta$  are used to classify whether the unlabeled sample is useful, uncertain, or useless. It should be noted that when the average normalized entropy of the two classifiers for an unlabeled sample is less than  $\delta$ , the unlabeled sample is considered useful; when the average normalized entropy is greater than  $\delta$  and less than  $\varepsilon$ , the sample is determined to be uncertain; when the average normalized entropy is greater than  $\varepsilon$ , the sample is considered useless. In the experiments, two types of classifiers, i.e., the K-nearest neighbor classifier with  $K = 3$  and the naive Bayes classifier, are used to evaluate the performance of the selected methods. Given a label rate  $\theta = 10\%$ , the results of the different methods are shown in Tables 5 and 6.

Table 5. Error rates of comparison methods on KNN classifier.

Dataset Name	ST-1V		ST-1VR		ST-2V		CT-2V		CT-TWD	
	Initial	Final	Initial	Final	Initial	Final	Initial	Final	Initial	Final
biode	0.3233	0.3357	0.3482	0.3459	0.3747	0.3705	0.3176	0.3067	0.2800	<b>0.2486</b>
cardio	0.3642	0.3722	0.3695	0.3588	0.3821	0.3722	0.2802	0.2889	0.2986	<b>0.2858</b>
cmc	0.3651	0.3830	0.3535	0.3671	0.4509	0.4510	0.3959	0.3731	0.3667	<b>0.3619</b>
frogs	0.4907	0.4936	0.4935	0.4913	0.5015	0.4996	0.4932	0.4932	0.4907	<b>0.4882</b>
hepatitis	0.2733	0.2667	0.2667	0.2667	0.3104	0.3000	0.2800	0.2600	0.3067	<b>0.2333</b>
hungarian	0.4493	0.4345	0.4441	0.4308	0.4204	0.4138	0.4276	0.4279	0.3759	<b>0.3448</b>
krvskp	0.3564	0.3395	0.3402	0.3561	0.3914	0.3843	0.3401	0.3226	0.3276	<b>0.3082</b>
lymph	0.3629	0.3500	0.3143	0.3288	0.3661	0.3557	0.3143	0.3143	0.2643	<b>0.2429</b>
newcylinder	0.4426	0.4463	0.4459	0.4422	0.4705	0.4622	0.4444	0.4630	0.4389	<b>0.4222</b>
pima	0.3237	0.3658	0.3278	0.3158	0.3846	0.3647	0.2974	0.2566	0.2408	<b>0.2395</b>
green	0.4333	0.4474	0.3852	0.3885	0.4725	0.4667	0.3702	0.3556	0.3486	<b>0.3345</b>
spectf	0.3374	0.3462	0.3061	0.3116	0.3812	0.3615	0.2862	0.2615	0.2538	<b>0.2346</b>
ttt	0.3326	0.3395	0.3421	<b>0.3326</b>	0.3468	0.3411	0.3642	0.3602	0.3539	0.3400
vehicle	0.3876	0.3977	0.3562	0.3482	0.3408	0.3369	0.2781	0.2764	0.2861	<b>0.2607</b>
vowel	0.3715	0.3485	0.3129	<b>0.3285</b>	0.3674	0.3581	0.3455	0.3512	0.3464	0.3452
wine	0.2624	0.2529	0.2354	0.2414	0.3248	0.2941	0.2476	0.2364	0.2353	<b>0.2059</b>
avg.	0.3673	0.3700	0.3526	0.3534	0.3929	0.3833	0.3427	0.3342	0.3259	<b>0.3060</b>

In Tables 5 and 6, the symbols “initial” and “final” denote the error rates of each model trained from labeled data and then improved by unlabeled data, respectively. All results in “initial” and “final” are obtained after averaging over 10-fold cross validation. In addition, for the convenience of comparison, the results with the lowest error rates are marked in bold. Tables 7 and 8 provide the computation time of different comparison methods on KNN and naive Bayes classifiers. The row “avg.” represents the average error rates of the selected models computed from all the datasets.

**Table 6.** Error rates of comparison methods on Naive Bayes classifier.

Dataset Name	ST-1V		ST-1VR		ST-2V		CT-2V		CT-TWD	
	Initial	Final	Initial	Final	Initial	Final	Initial	Final	Initial	Final
biode	0.3833	0.3895	0.3494	0.3401	0.3638	0.3529	0.3376	0.3386	0.3257	<b>0.3048</b>
cardio	0.2346	0.2369	0.2806	0.2792	0.2347	0.2311	0.2802	0.2797	0.2594	<b>0.2264</b>
cmc	0.3960	0.3953	0.3621	<b>0.3614</b>	0.4275	0.4252	0.3959	0.4007	0.3809	0.3741
frogs	0.4962	0.5022	0.4951	0.4912	0.4931	0.4925	0.4932	0.4844	0.4826	<b>0.4826</b>
hepatitis	0.3740	0.3767	0.3769	0.3627	0.3862	0.3767	0.3815	0.3684	0.3725	<b>0.3584</b>
hungarian	0.4910	0.4993	0.4655	0.4582	0.4884	0.4876	0.4676	0.4566	0.4585	<b>0.4483</b>
krvskp	0.2831	0.2897	0.2884	0.2859	0.2944	0.2884	0.3001	0.2801	0.2727	<b>0.2705</b>
lymph	0.2843	0.2986	0.3142	0.3075	0.3143	0.3129	0.3143	0.2943	0.2857	<b>0.2762</b>
newcylinder	0.4941	0.4952	0.4918	0.4992	0.5074	0.4970	0.4644	0.4778	0.4506	<b>0.4424</b>
pima	0.3266	0.3263	0.2672	0.2506	0.3332	0.3355	0.2674	0.2539	0.2408	<b>0.2368</b>
green	0.3109	0.3111	0.2783	0.2670	0.3389	0.3222	0.2402	0.2267	0.2283	<b>0.2186</b>
spectf	0.3246	0.3230	0.3195	0.3195	0.2992	0.2808	0.2862	0.2969	0.2476	<b>0.2308</b>
ttt	0.3322	0.3368	0.3200	0.3284	0.3474	0.3339	0.3342	0.3216	0.3206	<b>0.3163</b>
vehicle	0.1886	0.1810	0.1875	0.1857	0.1860	0.1831	0.1981	0.1852	0.1766	<b>0.1667</b>
vowel	0.1131	0.0939	0.1028	0.0912	0.1129	0.1051	0.1112	0.1020	0.1075	<b>0.0909</b>
wine	0.3018	0.3118	0.3056	0.3021	0.3053	0.3041	0.3076	0.2976	0.3042	<b>0.2941</b>
avg.	0.3334	0.3355	0.3253	0.3206	0.3395	0.3331	0.3237	0.3165	0.3071	<b>0.2961</b>

By observing Tables 5–8, it can be found that when the label rate is 10%, the initial performance of the ST-1VR model is better than ST-1V, and even on some datasets, such as “ttt” (33.26%), “vowel” (32.85%) in Table 5, and “cmc” (36.14%) in Table 6, it is better than that of the proposed model CT-TWD in this study, which shows the effectiveness of attribute reduction for semi-supervised learning. However, the improvement of both ST-1VR and ST-1V after learning unlabeled samples is not significant, even worse performance is observed on many datasets. The two-view self-training (ST-2V) can learn useful information from unlabeled samples and outperform the first two models in terms of performance. Combining Tables 7 and 8, it can be found that the computation time of ST-2V is greater than that of ST-1VR and ST-1V, which proves that two views have better performance than single views, but additional computational time is required to process them. For most datasets, the classifier retrained on unlabeled samples performs better than the classifier trained on labeled data only, while the co-training model with two views (CT-2V) achieved better performance using the KNN classifier and the naive Bayes classifier, which is improved by 2.50% and 2.20%, respectively, because of the mutual learning in the two classifiers. The average error rates on the KNN classifier and the naive Bayes classifier are lower than ST-1V, ST-1VR, and ST-2V, which demonstrates the stability of CT-2V. In addition, CT-2V requires to simultaneously train the two classifiers, resulting in a slightly longer computation time. However, the results in Tables 5 and 6 show that the performance of CT-2V is with the average error rates of 33.42% on the KNN classifier and 31.65% on the naive Bayes classifier, which still has a large gap compared to the proposed model CT-TWD, with 30.60% on the KNN classifier and 29.61% on the naive Bayes classifier. In terms of the calculation time, although the average calculation time (avg.) of CT-TWD in Tables 7 and 8 is relatively large with 4.0440 s on the KNN classifier and 2.5763 s on the naive Bayes classifier, considering the good performance of CT-TWD, the additional time cost is clearly acceptable.

To compare the differences among the methods more comprehensively, we also conduct experiments at different label rates, and the results are shown in Figures 2 and 3.

As can be seen in Figures 2 and 3, the proposed model CT-TWD can learn from unlabeled samples and achieve impressive performance against different models. ST-1V is a single-view semi-supervised learning model, and it can be found in the experiments that ST-1V performs poorly on most datasets; even worse performance occurs at higher label rates, such as “lymph” with the KNN classifier and “frogs” with the naive Bayes classifier. This may be because the initially labeled data are not representative, so the classifiers will mislabel

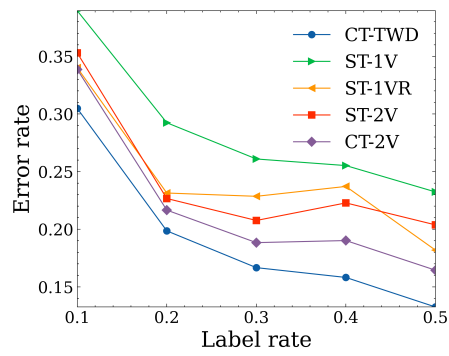
unlabeled samples in the training process. Therefore, the classifier will learn the wrong classification information, which results in poor generalization of the final performance. ST-1VR is also a single-view semi-supervised self-learning model but performs attribute reduction on the dataset. Although its overall performance is poor, it outperforms ST-1V, which shows the effectiveness of the semi-supervised neighborhood discriminant index-based attribute reduction method. However, ST-1VR still has poor final performance with the limitations of the single-view model, such as “cmc” and “lymph” with the KNN classifier. ST-2V is a multi-view self-training model that uses randomly split subsets of attributes from the raw dataset to train the base classifiers, and a threshold is used to select useful samples to help the classifiers retrain themselves, but its final performance is not good. On the one hand, the two classifiers of ST-2V are self-taught. On the other hand, the poor quality of the randomly partitioned subsets of attributes also leads to the disappointing performance of ST-2V. Although CT-2V can make two base classifiers learn from each other through unlabeled samples to improve the performance, the two subspaces of CT-2V are randomly divided from the dataset. Therefore, the performance of the classifiers is not stable, resulting in CT-2V only performing better on some datasets.

**Table 7.** Computation time of comparison methods on KNN classifier (in seconds).

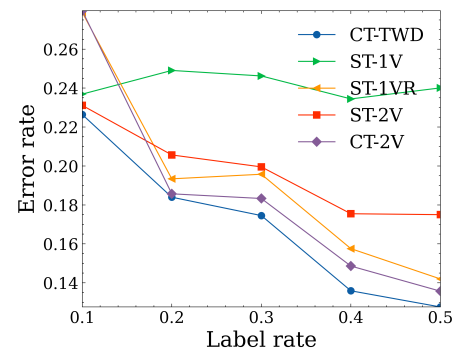
Dataset Name	ST-1V	ST-1VR	ST-2V	CT-2V	CT-TWD
biode	0.7241	0.8591	0.8357	1.2159	1.7486
cardio	1.4974	1.7692	2.7157	3.2531	4.2960
cmc	1.1913	1.0838	1.2039	1.1448	1.9738
frogs	8.5527	15.1085	18.9194	20.2269	20.6095
hepatitis	0.0837	0.0947	0.1478	0.1771	0.2135
hungarian	0.1721	0.1678	0.2480	0.3563	0.4245
krvskp	2.4541	6.8275	2.7247	4.5600	5.8669
lymph	0.0764	0.0973	0.1131	0.1320	0.1882
newcylinder	0.2567	0.2988	0.3335	0.4090	0.4505
pima	0.5108	0.6147	0.6824	0.6968	0.6937
green	0.0522	0.0566	0.0717	0.0811	0.0859
spectf	0.1430	0.2100	0.1855	0.2539	0.2629
ttt	0.7839	0.7737	0.9943	0.9106	0.9787
vehicle	0.5464	0.6936	1.1177	1.1366	1.1564
vowel	1.0919	1.0220	1.4024	1.1234	1.3427
wine	0.1033	0.1125	0.1553	0.1220	0.1485
avg.	1.1400	1.8619	3.1851	3.5800	4.0440

**Table 8.** Computation time of comparison methods on Naive Bayes classifier (in seconds).

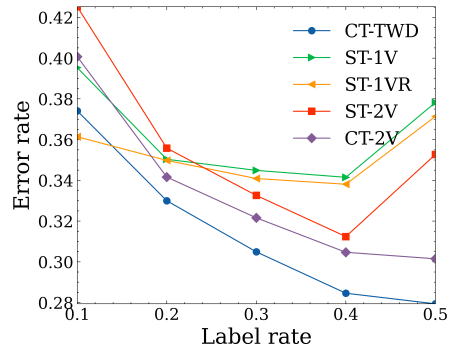
Dataset Name	ST-1V	ST-1VR	ST-2V	CT-2V	CT-TWD
biode	0.6475	0.6839	0.8281	0.9782	1.0773
cardio	1.3194	1.3785	1.5004	1.6889	1.7402
cmc	0.7274	0.7704	0.9232	0.9902	1.0686
frogs	6.1380	9.8430	11.2920	14.1749	14.9361
hepatitis	0.0685	0.0748	0.0864	0.1038	0.1090
hungarian	0.1281	0.1442	0.1454	0.2429	0.2635
krvskp	2.9206	2.8730	2.4105	2.4737	2.6089
lymph	0.0804	0.0714	0.0742	0.0844	0.0954
newcylinder	0.2323	0.2806	0.3145	0.3498	0.3535
pima	0.3683	0.3794	0.4796	0.6669	0.6978
green	0.0397	0.0492	0.0618	0.0694	0.0752
spectf	0.1349	0.1382	0.1986	0.2465	0.2790
ttt	0.4538	0.4345	0.7001	0.6771	0.7709
vehicle	0.4709	0.4555	0.7014	0.7113	0.8108
vowel	0.5423	0.5481	0.5310	0.6838	0.7396
wine	0.0760	0.0830	0.1128	0.1166	0.1374
avg.	0.8967	1.1380	1.2725	2.4258	2.5763



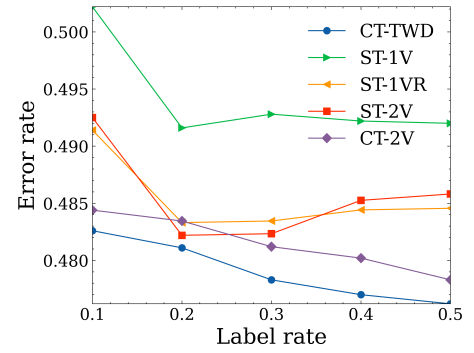
(a) biode



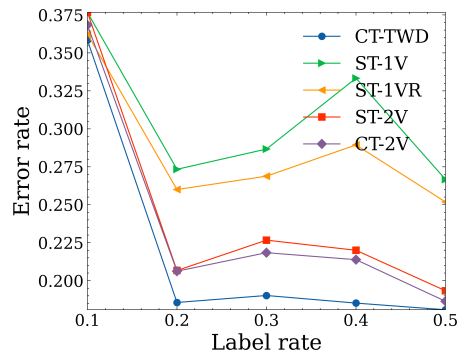
(b) cardio



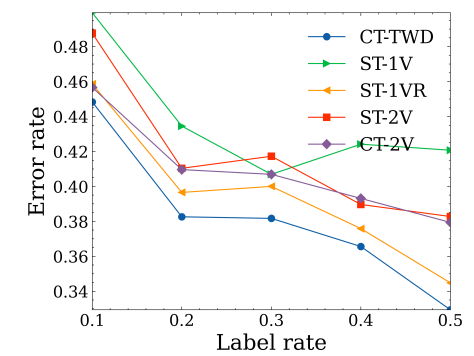
(c) cmc



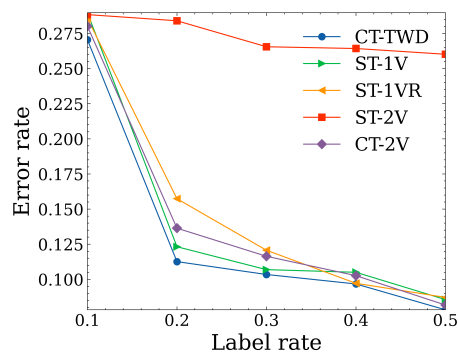
(d) frogs



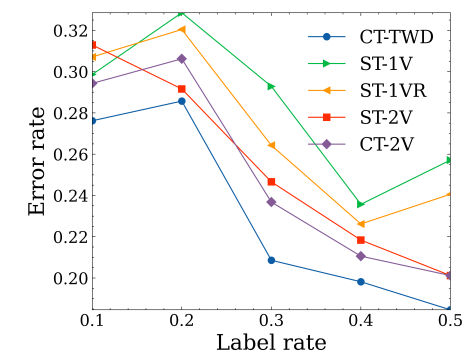
(e) hepatitis



(f) hungarian



(g) krvskp



(h) lymph

Figure 2. Cont.

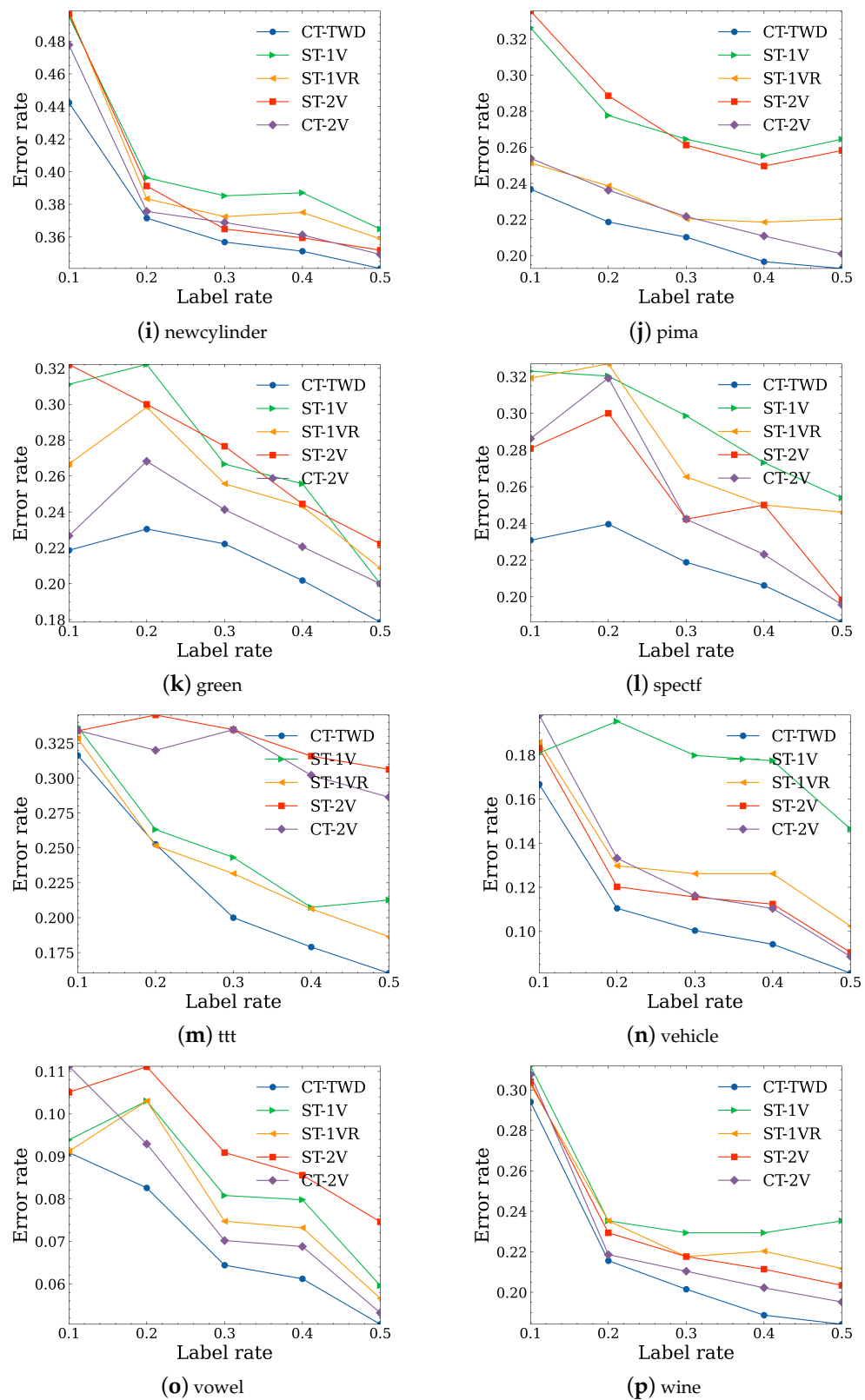
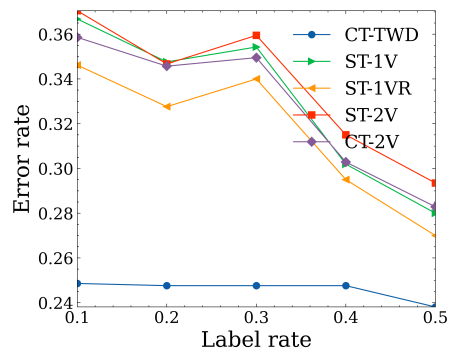
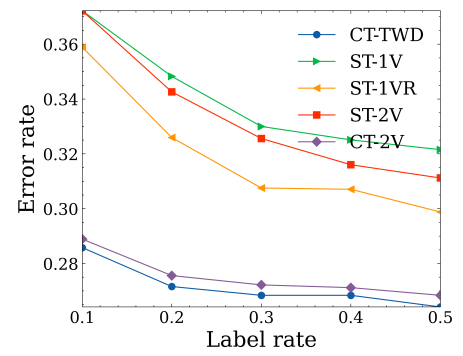


Figure 2. Error rates of comparison methods under different label rates when using KNN.

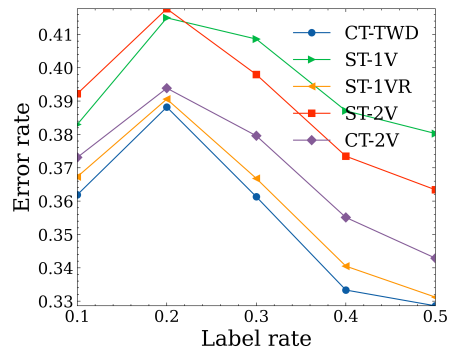




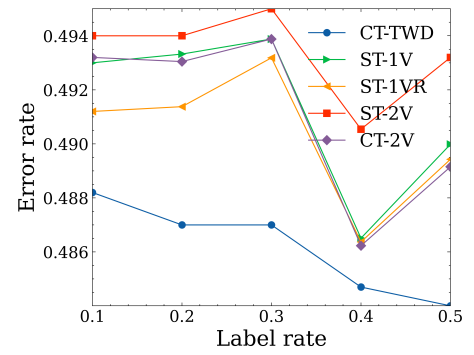
(a) biode



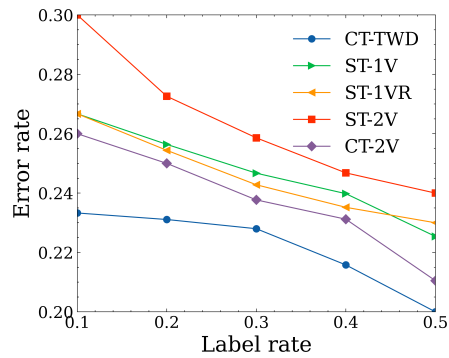
(b) cardio



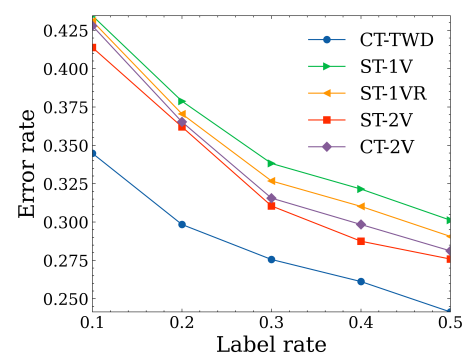
(c) cmc



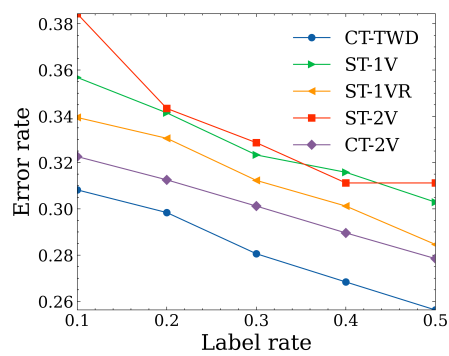
(d) frogs



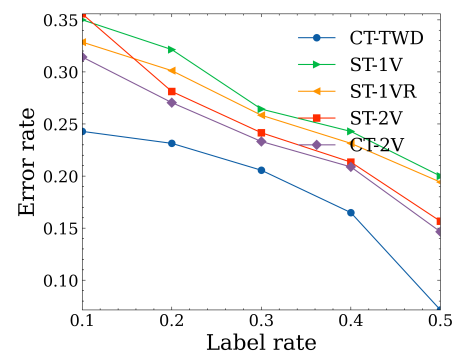
(e) hepatitis



(f) hungarian

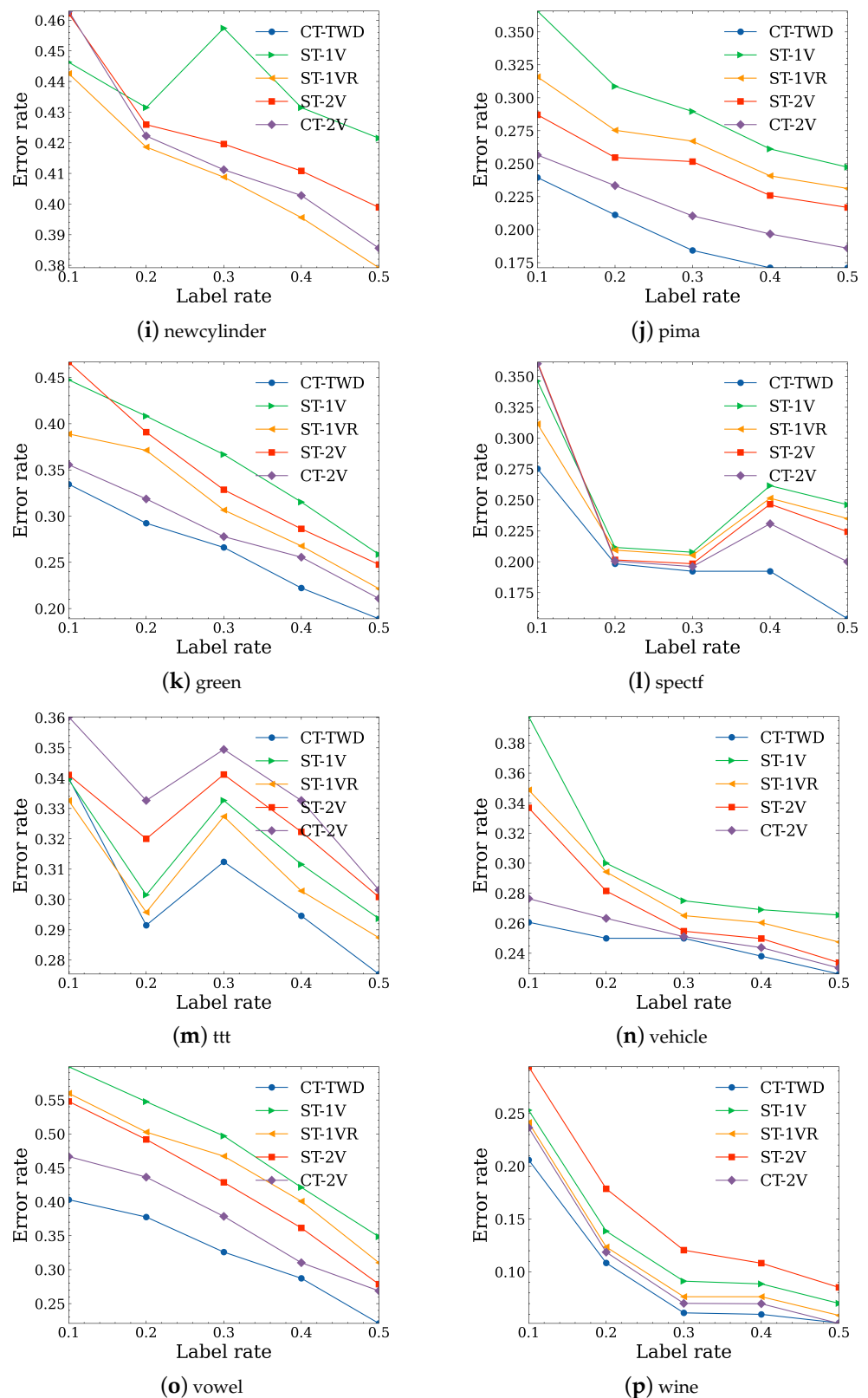


(g) krvskp



(h) lymph

Figure 3. Cont.



**Figure 3.** Error rates of comparison methods under different label rates when using Naive Bayes.

Different from the selected comparison models, the CT-TWD uses the three-way co-decision model in the training process to classify unlabeled samples into useful, uncertain, and useless. The training set of each classifier is updated only when the unlabeled samples are useful and have a positive impact on the model performance. Such a sample selec-

tion mechanism ensures that CT-TWD can effectively use unlabeled samples to improve performance on most datasets at different label rates. For example, the proposed model achieves an improvement of 22.03% at a 30% label rate on the “vowel” dataset with the KNN classifier and an improvement of 13.55% at a 40% label rate on the “wine” dataset with the naive Bayes classifier, illustrating the potential of the proposed model for partially labeled data.

It should be noted that for some datasets, such as “cmc” on the naive Bayes classifier and “lymph” on the KNN classifier, as the label rate increases, the performance of the methods tends to decrease. This is likely due to the labeled data not being representative enough, thereby limiting the performance of the classifier as the data scale increases. Compared to other models, the CT-TWD proposed in this study assigns pseudo labels of 0 and 1 to unlabeled samples to form two views of data, which makes the two views of data still maintain the discriminative ability as the raw dataset. Therefore, the quality of the base classifiers obtained by CT-TWD has good robustness, which allows it to have better performance across all the datasets.

## 5. Conclusions

In real-world applications, annotating large amounts of data is often challenging, but collecting unlabeled data is relatively easy, which results in semi-supervised data with a small amount of labeled data and a large amount of unlabeled data. In this study, we proposed a simple yet effective strategy for generating pseudo-labels for partially labeled data and developed a heuristic semi-supervised attribute reduction algorithm using a granular conditional entropy measure. To exploit useful unlabeled samples for learning, we combined the three-way decision with the normalized entropy and proposed a three-way co-decision model for partially labeled data. However, due to the proposed model requiring partitioning of the dataset with 0 and 1 pseudo labels, which can only have significant advantages in binary classification problems, it still has limitations in multi-classification problems. Therefore, extending the model to multi-classification problems will be future work. Also, exploring the semi-supervised model for discrete and continuous data is worthy of further investigation.

**Author Contributions:** Conceptualization, C.G.; Methodology, L.W. and C.G.; Software, L.W., J.Z. and J.W.; Data analysis, L.W.; Writing the original draft, L.W.; Review and editing, C.G., J.Z. and J.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the Shenzhen Science and Technology Program (No. JCYJ20210324094601005), the Natural Science Foundation of Guangdong Province, China (No. 2021A1515011861), the National Natural Science Foundation of China (Nos. 62076164 and 61806127), and Shenzhen Institute of Artificial Intelligence and Robotics for Society.

**Data Availability Statement:** Data are available from the corresponding author upon reasonable request.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Li, J.D.; Cheng, K.W.; Wang, S.H.; Morstatter, F.; Trevino, R.P.; Tang, J.L.; Liu, H. Feature Selection: A Data Perspective. *ACM Comput. Surv.* **2017**, *50*, 94.
2. Thangavel, K.; Pethalakshmi, A. Dimensionality reduction based on rough set theory: A review. *Appl. Soft Comput.* **2009**, *9*, 1–12. [[CrossRef](#)]
3. Pawlak, Z. *Rough Sets: Theoretical Aspects of Reasoning about Data*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 1991.
4. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356. [[CrossRef](#)]
5. Hu, M.; Tsang, E.C.; Guo, Y.; Chen, D.; Xu, W. A novel approach to attribute reduction based on weighted neighborhood rough sets. *Knowl.-Based Syst.* **2021**, *220*, 106908. [[CrossRef](#)]
6. Xu, W.H.; Li, W.T. Granular computing approach to two-way learning based on formal concept analysis in fuzzy datasets. *IEEE Trans. Cybern.* **2014**, *46*, 366–379. [[CrossRef](#)]
7. Zhang, P.; Li, T.; Wang, G.; Luo, C.; Chen, H.; Zhang, J.; Wang, D.; Yu, Z. Multi-source information fusion based on rough set theory: A review. *Inf. Fusion.* **2021**, *68*, 85–117. [[CrossRef](#)]

8. Pawlak, Z.; Wong, S.K.M.; Ziarko, W. Rough sets: Probabilistic versus deterministic approach. *Int. J. Man-Mach. Stud.* **1988**, *29*, 81–95. [[CrossRef](#)]
9. Sun, L.; Wang, L.Y.; Ding, W.P.; Qian, Y.H.; Xu, J.C. Neighborhood multi-granulation rough sets-based attribute reduction using Lebesgue and entropy measures in incomplete neighborhood decision systems. *Knowl.-Based Syst.* **2020**, *192*, 105373. [[CrossRef](#)]
10. Jiang, F.; Sui, Y.S.; Zhou, L. A relative decision entropy-based feature selection approach. *Pattern Recognit.* **2015**, *48*, 2151–2163. [[CrossRef](#)]
11. Gao, C.; Lai, Z.H.; Zhou, J.; Wen, J.J.; Wong, W.K. Granular maximum decision entropy-based monotonic uncertainty measure for attribute reduction. *Int. J. Approx. Reason.* **2019**, *104*, 9–24. [[CrossRef](#)]
12. Gao, C.; Lai, Z.H.; Zhou, J.; Zhao, C.R.; Miao, D.Q. Maximum decision entropy-based attribute reduction in decision-theoretic rough set model. *Knowl.-Based Syst.* **2018**, *143*, 179–191. [[CrossRef](#)]
13. Yang, X.; Liang, S.; Yu, H.; Gao, S.; Qian, Y. Pseudo-label neighborhood rough set: Measures and attribute reductions. *Int. J. Approx. Reason.* **2019**, *105*, 112–129. [[CrossRef](#)]
14. Yuan, Z.; Chen, H.M.; Li, T.R. Exploring interactive attribute reduction via fuzzy complementary entropy for unlabeled mixed data. *Pattern Recognit.* **2022**, *127*, 108651. [[CrossRef](#)]
15. Xu, W.H.; Li, M.M.; Wang, X.Z. Information fusion based on information entropy in fuzzy multi-source incomplete information system. *Int. J. Fuzzy Syst.* **2017**, *19*, 1200–1216. [[CrossRef](#)]
16. Liang, D.C.; Cao, W.; Xu, Z.S.; Wang, M.W. A novel approach of two-stage three-way co-opetition decision for crowdsourcing task allocation scheme. *Inf. Sci.* **2021**, *559*, 191–211. [[CrossRef](#)]
17. Qian, J.; Liu, C.H.; Miao, D.Q.; Yue, X.D. Sequential three-way decisions via multi-granularity. *Inf. Sci.* **2020**, *507*, 606–629. [[CrossRef](#)]
18. Xu, W.H.; Guo, Y.T. Generalized multigranulation double-quantitative decision-theoretic rough set. *Knowl.-Based Syst.* **2016**, *105*, 190–205. [[CrossRef](#)]
19. Yang, J.L.; Yao, Y.Y. A three-way decision based construction of shadowed sets from Atanassov intuitionistic fuzzy sets. *Inf. Sci.* **2021**, *577*, 1–21. [[CrossRef](#)]
20. Yao, Y.Y. Three-way granular computing, rough sets, and formal concept analysis. *Int. J. Approx. Reason.* **2020**, *116*, 106–125. [[CrossRef](#)]
21. Yao, Y.Y. Tri-level thinking: Models of three-way decision. *Int. J. Mach. Learn. Cybern.* **2020**, *11*, 947–959. [[CrossRef](#)]
22. Yao, Y.Y. Three-way decisions with probabilistic rough sets. *Inf. Sci.* **2010**, *180*, 341–353. [[CrossRef](#)]
23. Yao, Y.Y. The superiority of three-way decisions in probabilistic rough set models. *Inf. Sci.* **2011**, *181*, 1080–1096. [[CrossRef](#)]
24. Yue, X.D.; Chen, Y.F.; Miao, D.Q.; Fujita, H. Fuzzy neighborhood covering for three-way classification. *Inf. Sci.* **2020**, *507*, 795–808. [[CrossRef](#)]
25. Yao, Y.Y. Three-way decision and granular computing. *Int. J. Approx. Reason.* **2018**, *103*, 107–123. [[CrossRef](#)]
26. Li, J.H.; Huang, C.C.; Qi, J.J.; Qian, J.H.; Liu, W.Q. Three-way cognitive concept learning via multi-granularity. *Inf. Sci.* **2017**, *378*, 244–263. [[CrossRef](#)]
27. Wang, G.Y.; Yu, H.; Li, T.R. Decision region distribution preservation reduction in decision-theoretic rough set model. *Inf. Sci.* **2014**, *278*, 614–640.
28. Ren, R.S.; Wei, L. The attribute reductions of three-way concept lattices. *Knowl.-Based Syst.* **2016**, *99*, 92–102. [[CrossRef](#)]
29. Huang, Q.Q.; Li, T.R.; Huang, Y.Y.; Yang, X. Incremental three-way neighborhood approach for dynamic incomplete hybrid data. *Inf. Sci.* **2020**, *541*, 98–122. [[CrossRef](#)]
30. Zhang, X.Y.; Zhou, Y.H.; Tang, Y.; Fan, Y.R. Three-way improved neighborhood entropies based on three-level granular structures. *Int. J. Mach. Learn. Cybern.* **2022**, *13*, 1861–1890. [[CrossRef](#)]
31. Kong, Q.Z.; Zhang, X.W.; Xu, W.H.; Long, B.H. A novel granular computing model based on three-way decision. *Int. J. Approx. Reason.* **2022**, *144*, 92–112. [[CrossRef](#)]
32. Fang, Y.; Gao, L.; Liu, Z.H.; Yang, X. Generalized cost-sensitive approximate attribute reduction based on three-way decisions. *J. Nanjing Univ. Sci. Technol.* **2019**, *43*, 481–488.
33. Miao, D.Q.; Gao, C.; Zhang, N.; Zhang, Z.F. Diverse reduct subspaces based co-training for partially labeled data. *Int. J. Approx. Reason.* **2011**, *52*, 1103–1117. [[CrossRef](#)]
34. Wang, R.; Chen, D.G.; Kwong, S. Fuzzy-rough-set-based active learning. *IEEE Trans. Fuzzy Syst.* **2013**, *22*, 1699–1704. [[CrossRef](#)]
35. Li, B.Y.; Xiao, J.M.; Wang, X.H. Feature selection for partially labeled data based on neighborhood granulation measures. *IEEE Access* **2019**, *7*, 37238–37250. [[CrossRef](#)]
36. Liu, K.Y.; Yang, X.B.; Yu, H.L.; Mi, J.S.; Wang, P.X.; Chen, X.J. Rough set based semi-supervised feature selection via ensemble selector. *Knowl.-Based Syst.* **2019**, *165*, 282–296. [[CrossRef](#)]
37. Pan, L.C.; Gao, C.; Zhou, J. Three-way decision-based tri-training with entropy minimization. *Inf. Sci.* **2022**, *610*, 33–51. [[CrossRef](#)]
38. Ash, R.B. *Information Theory*; Courier Corporation: Chelmsford, MA, USA, 2012.
39. Ashby, D.; Smith, A.F.M. Evidence-based medicine as Bayesian decision-making. *Stat. Med.* **2000**, *19*, 3291–3305. [[CrossRef](#)]
40. Blum, A.; Mitchell, T. Combining labeled and unlabeled data with co-training. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, Madison, WI, USA, 24–26 July 1998; pp. 92–100.
41. Dai, D.; Li, H.X.; Jia, X.Y.; Zhou, X.Z.; Huang, B.; Liang, S.N. A co-training approach for sequential three-way decisions. *Int. J. Mach. Learn. Cybern.* **2020**, *11*, 1129–1139. [[CrossRef](#)]

42. Wang, G.Y.; Yu, H.; Yang, D.C. Decision table reduction based on conditional information entropy. *Chin. J. Comput.* **2002**, *25*, 759–766.
43. Zhu, X.J.; Goldberg, A.B. *Introduction to Semi-Supervised Learning*; Morgan and Claypool Publishers: Cambridge, MA, USA, 2009.
44. Liang, J.Y.; Shi, Z.Z. The information entropy, rough entropy and knowledge granulation in rough set theory. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **2004**, *12*, 37–46. [[CrossRef](#)]
45. Gao, C.; Zhou, J.; Miao, D.; Wen, J.J.; Yue, X.D. Three-way decision with co-training for partially labeled data. *Inf. Sci.* **2021**, *544*, 500–518. [[CrossRef](#)]
46. Witten, I.H.; Frank, E. Data mining: Practical machine learning tools and techniques with Java implementations. *ACM Sigm. Rec.* **2002**, *31*, 76–77. [[CrossRef](#)]
47. Nigam, K.; Ghani, R. Analyzing the effectiveness and applicability of co-training. In Proceedings of the Ninth International Conference on Information and Knowledge Management, McLean, VA, USA, 6–11 November 2000; pp. 86–93.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.