Applied Intelligence (2023) 53:17708–17726
https://doi.org/10.1007/s10489-022-04405-3

Rough sets-based tri-trade for partially labeled data

Ziming Luo[1,2] · **Can Gao**[1,2,3] · **Jie Zhou**[1,2,3]

Accepted: 13 December 2022 / Published online: 11 January 2023

Abstract
The theory of rough sets is one of the most representative models for handling supervised data entangled with vagueness, impreciseness, or uncertainty. However, little work has been devoted to learning from partially labeled data using rough sets. In this study, a rough sets-based tri-trade model is proposed for partially labeled data. More specifically, a new discernibility matrix that considers both labeled and unlabeled data is first proposed, based on which a beam search-based heuristic algorithm is provided to generate multiple semi-supervised reducts. Then, a tri-trade model using three diverse semi-supervised reducts is developed, in which a data editing technique is embedded to generate reliable pseudo-labels for unlabeled data to improve the tri-trade model. Both theoretical analysis and comparative experiments on the UCI datasets show that the proposed model can effectively utilize unlabeled data to improve generalization performance and compare favorably to other representative methods.

Keywords Rough sets · Attribute reduction · Discernibility matrix · Tri-trade · Partially labeled data

1 Introduction

Rough set theory [1] is an effective method for dealing with vague, imprecise, or uncertain data and has been widely used in various fields such as machine learning, pattern recognition, and data mining [2–5]. In rough set theory, each attribute or subset of attributes is considered to be an indiscernibility relation. On this basis, a rough set is defined by two approximations, called the lower and upper approximations, to represent vague, imprecise, or uncertain concepts [6]. Attribute reduction is a primary research topic in rough sets [7–19]. It aims to remove irrelevant and redundant attributes while retaining important attributes to maintain discriminating power over the data. Many rough sets-based attribute reduction methods have been proposed, including positive region-based, discernibility matrix-based, and information entropy-based methods. Among them, discernibility matrix-based methods have received extensive attention because of their simplicity and ease of implementation [20–24], where heuristic methods initialized with core attributes are often used to generate optimal reducts.

Existing rough sets-based attribute reduction algorithms mainly deal with labeled data or unlabeled data. However, in many practical tasks, such as web-page categorization [25], intrusion detection [26], and medical diagnosis [27], unlabeled training objects are easily available, but labeled ones are difficult to obtain because labeling objects is labor intensive and expensive. If only a small number of labeled objects are used, it often causes overfitting of the model to the data, resulting in weak generalization ability. Additionally, learning a supervised model without considering unlabeled objects leads to a massive waste of exploitable data. Thus, how to utilize a large number of unlabeled objects to improve learning performance has emerged as a hot research topic in machine learning.

In the rough sets field, many scholars have researched the topic of attribute reduction based on discernibility matrix. Wei et al. [20] developed a discernibility-matrix based incremental attribute reduction algorithm to generate the optimal reduct of dynamic data. Ma et al.

✉ Can Gao
2005gaocan@163.com

[1] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, 518060, People's Republic of China

[2] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen, 518060, People's Republic of China

[3] SZU Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, 518060, People's Republic of China

[21] constructed a compressed binary discernibility matrix for group dynamic data and developed an incremental attribute reduction algorithm, which considered both single dynamic objects and group dynamic objects. By optimizing the space constraint of storing discernibility matrix, Liu et al. [22] designed an incremental attribute method on fused decision table. Additionally, some attribute reduction methods have been proposed for partially labeled data. Dai et al. [28] introduced the concept of discernibility pair and developed two attribute reduction measures for partially labeled categorical data. Based on mutual information, Hu et al. [29] defined the significance measure for attributes in partially labeled data and utilized it as heuristic information to speed up the attribute reduction process. Xie et al. [30] proposed two types of induced hypergraphs for partially labeled decision systems and designed a fast algorithm based on low-complexity heuristics to compute the optimal reduct. Unlike traditional methods that use only one fitness function, Liu et al. [31] introduced an ensemble voting mechanism to select a more appropriate semi-supervised reduct by constructing multiple fitness functions. Gao et al. [32] generated proxy labels for unlabeled data using prior class-distribution information and developed the granular conditional entropy measure for semi-supervised attribute reduction. In addition, some related semi-supervised learning methods have also been proposed. Wang et al. [33] used Gaussian kernel-based fuzzy rough sets to measure the inconsistency of unlabeled objects and provided an active learning model based on SVM. By integrating three-way decision theory and cost-sensitive learning, Min et al. [34] developed an active learning model based on the k-nearest neighbor classifier. By introducing the idea of tri-partition in the three-way decision, Gao et al. [23] proposed the three-way co-decision model to improve the semi-supervised learning performance. In addition, rough set theory has also been successfully applied to address some practical semi-supervised tasks, such as short text classification [35], defect detection [36], relationship categorization [37, 38], and so on [39].

The aforementioned studies primarily focus on rough sets-based semi-supervised attribute reduction or practical applications. However, little work has been devoted to the construction of semi-supervised models to directly learn from partially labeled data using rough sets. Tri-training [40] is a typical disagreement-based semi-supervised model that employs three learners to learn from each other using unlabeled data but encounters the problems of the weak diversity of the base learners and low quality of selected unlabeled data. In this study, we propose a rough sets-based tri-trade model for partially labeled data. The primary contributions are as follows:

1. To address the attribute reduction problem for partially labeled data, a new semi-supervised discernibility matrix is proposed, based on which a beam search-based heuristic attribute reduction algorithm is designed to generate optimal semi-supervised reducts. The semi-supervised discernibility matrix considers both labeled and unlabeled data and allows for a certain degree of inconsistency, which contributes to improving the robustness and adaptability of semi-supervised attribute reduction.
2. To learn from unlabeled data, a tri-trade model that uses three diverse semi-supervised reducts to train base classifiers is constructed, and a novel data editing technique is developed to reliably identify useful unlabeled data. By selecting useful unlabeled objects and simultaneously eliminating mislabeled unlabeled objects, the proposed data editing technique can enable the base classifiers to learn from each other on high-quality unlabeled data.
3. To obtain insight into the proposed model, a theoretical analysis is offered from the perspective of noise learning. Furthermore, extensive experiments are carried out to validate the effectiveness of the proposed model, and good results are achieved.

The rest of this paper is organized as follows. Section 2 introduces some concepts in rough sets and semi-supervised learning, respectively. Section 3 provides a detailed description of the proposed tri-trade model for partially labeled data as well as the theoretical analysis. Section 4 reports the experimental results and analysis. Finally, Section 5 summarizes the paper.

2 Preliminaries

In this section, some concepts related to rough sets and semi-supervised learning are briefly reviewed. A detailed description of these theories can be found in [1, 6, 41–45].

2.1 Rough sets

In rough set theory, data of interest can be represented in an information system [6]. An information system consists of quadruple, denoted as $IS = (U, A, V, f)$, where U is a non-empty finite set of objects, called the universe; A is a non-empty finite set of attributes; V is the union of attribute domains, and let V_a denote the domain of attribute a such that $V = \bigcup V_a$ for each $a \in A$; f is called the information function such that $f(x, a) \in V_a$ for each $x \in U$ and $a \in A$, which assigns a unique value to each attribute of an object in U. When the attribute set A can be categorized into condition attribute set C and decision attribute set D

and $C \cap D = \emptyset$, the information system is also called a decision table or decision information system [6].

Definition 1 Let $IS = (U, A = C \cup D, V, f)$ be a decision table. For any non-empty attribute subset $B \subseteq A$, the indiscernibility relation induced by B is defined as:

$$IND(B) = \{(x, y) \in U \times U : a(x) = a(y), \forall a \in B\} \quad (1)$$

Definition 2 Let $IS = (U, A = C \cup D, V, f)$ be a decision table and $IND(B)$ be the equivalence relation induced by an attribute subset $B \subseteq A$, the set of equivalence classes of U induced by $IND(B)$ is denoted as:

$$U/IND(B) = \bigcup \{[x]_B : x \in U\}, \quad (2)$$

where $[x]_B = \{y \in U : (x, y) \in IND(B)\}$ is called the equivalence class of x under the equivalence relation $IND(B)$.

Definition 3 Let $IS = (U, A = C \cup D, V, f)$ be a decision table. For any subset X of U , the lower and upper approximations with respect to an attribute subset $B \subseteq A$ are defined as:

$$\begin{aligned} \underline{B}(X) &= \{x \in U \mid [x]_B \subseteq X\} \\ \overline{B}(X) &= \{x \in U \mid [x]_B \cap X = \emptyset\} \end{aligned} \quad (3)$$

$\underline{B}(X)$ is also called the B-positive region of X over U, denoted as $POS_B(X)$. The difference set between $\overline{B}(X)$ and $\underline{B}(X)$ is called the B-boundary region of X over U, denoted as $BND_B(X)$. And the universe after removing the objects in $\overline{B}(X)$ is called the B-negative region of X over U, denoted as $NEG_B(X)$, namely $NEG_B(X) = U - \overline{B}(X)$.

Definition 4 Let $IS = (U, A = C \cup D, V, f)$ be a decision table and $U/D = \{Y_1, Y_2, \ldots, Y_{|U/D|}\}$ be the partition derived from the decision attribute D over U. The positive, boundary, and negative regions of D given an attribute subset $B \subseteq C$ are defined as:

$$\begin{aligned} POS_B(D) &= \bigcup_{Y_i \in U/D} \underline{B}(Y_i) \\ BND_B(D) &= \bigcup_{Y_i \in U/D} \left(\overline{B}(Y_i) - \underline{B}(Y_i)\right) \\ NEG_B(D) &= U - \bigcup_{Y_i \in U/D} \overline{B}(Y_i) \end{aligned} \quad (4)$$

Definition 5 Let $IS = (U, A = C \cup D, V, f)$ be a decision table. For an attribute subset S of C, S is a reduct of C if and only if:

1) $POS_S(D) = POS_C(D)$ and
2) $\forall a \in S, POS_{S-\{a\}}(D) \neq POS_S(D)$

Meanwhile, the classification ability of an attribute or attribute subset can be represented by a discernibility matrix whose entry describes the discernible information to each pair of objects with different decisions. Formally, the disernibility matrix, the core attribute, and its attribute reduction are defined as follows.

Definition 6 Let $IS = (U, A = C \cup D, V, f)$ be a decision table. The element of the discernibility matrix M is denoted as:

$$e_{ij} = \begin{cases} \left\{a \in C \mid a(x_i) \neq a\left(x_j\right)\right\}, & d(x_i) \neq d\left(x_j\right) \\ \varnothing, & \text{otherwise} \end{cases} \quad (5)$$

In the discernibility matrix, if two objects have different decisions, the element is a set of discernable attributes, on each of which the two objects have different values; otherwise, the element is empty.

Definition 7 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 $e = \{a\}$.

Definition 8 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 S of C, S is a reduct of C if and only if:

1. $\forall e \in M \wedge e \neq \emptyset, S \cap e \neq \emptyset$ and
2. $\forall a \in S \wedge S^* = S - \{a\}, \exists e \in M \wedge S^* \cap e = \emptyset$

Different from the positive region-based reduct in Definition 5, the reduct in Definition 8 is a minimal subset of attributes that intersects with every non-empty element in the discernibility matrix. In other words, a discernibility matrix-based reduct is a jointly sufficient and individually necessary attribute subset to discriminate all objects in the original data.

2.2 Semi-supervised learning

In semi-supervised learning, a given partially labeled data $U = L \cup N$ contains a set of labeled objects $L = \{x_i, y_i\}_{i=1}^{l}$ and a set of unlabeled objects $N = \left\{x_j, ?\right\}_{j=l+1}^{l+n}$, where x_i and x_j are described by m attributes, y_i belongs to one of the k classes or unknown, and $l \ll n$. Generally, semi-supervised learning can be classified as semi-supervised clustering, semi-supervised classification, and semi-supervised regression [42, 46, 47]. This paper mainly concentrates on semi-supervised classification.

Semi-supervised classification aims to exploit a large amount of unlabeled data to improve the learner trained only on labeled data. Semi-supervised classification methods can be roughly classified as low-density separation methods, generative methods, graph-based methods, and disagreement-based methods [47]. Self-training [48] is a classic semi-supervised method that retrains the learner by self-labeled objects. More specifically, the model first trains a classifier on labeled objects, and then iteratively annotates some confidently unlabeled objects to retrain the classifier. Co-training [23] is a disagreement-based model that could enable two classifiers to learn from each other on unlabeled data. Standard co-training requires two sufficient and redundant views to describe data. On each view, a classifier is trained on labeled objects, and then the two classifiers share some unlabeled objects with high confidence prediction labels to improve each other. Tri-training [40] is another popular disagreement-based model. It resamples the set of labeled objects to obtain three labeled training sets, on each of which a base classifier is trained. In each iteration of tri-training, if two classifiers have the same prediction on an unlabeled object, this object with its predicted label is used to update the third classifier, until the stopping condition is met.

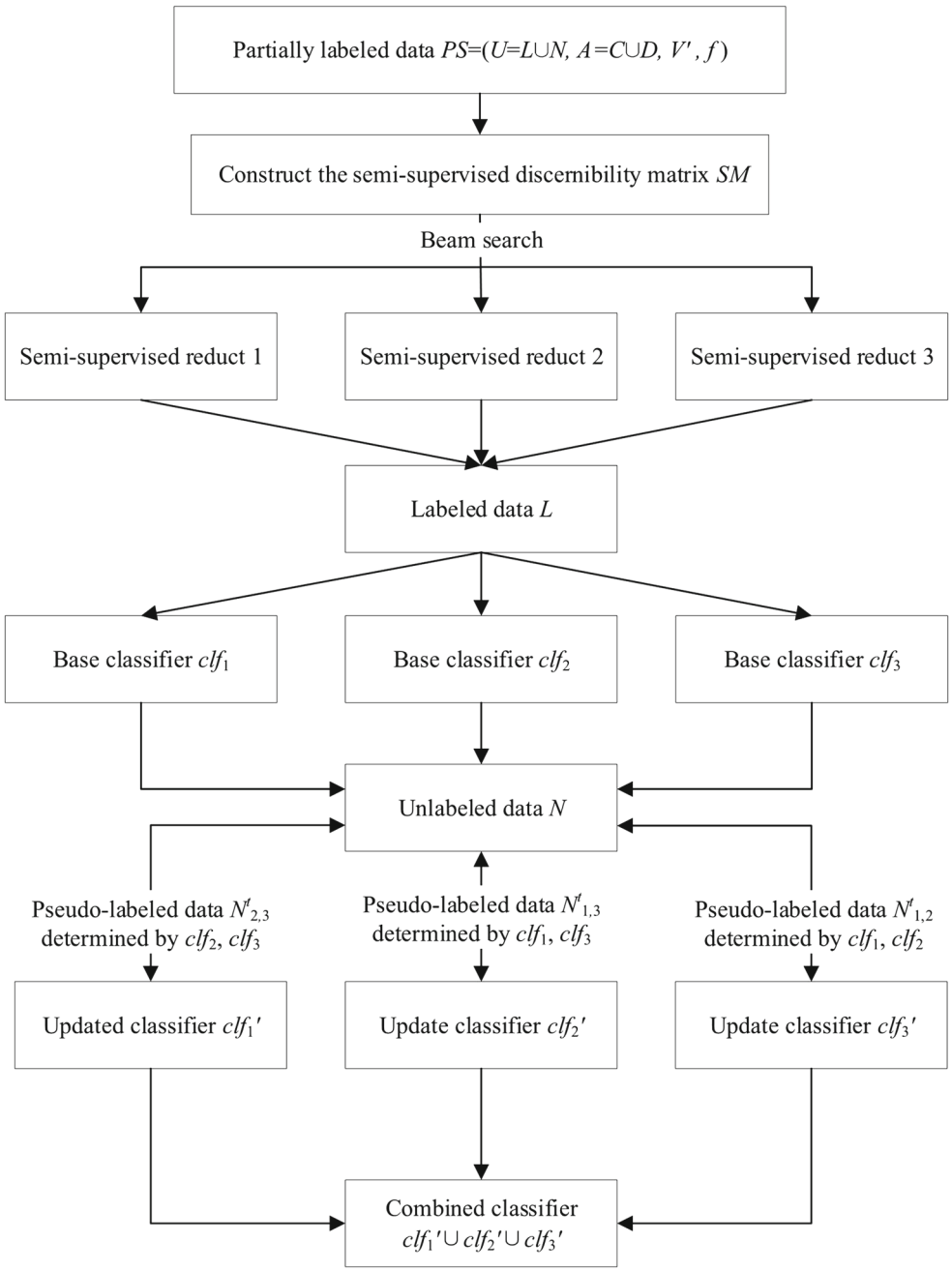


Fig. 1 Framework of the proposed tri-trade model

However, standard tri-training suffers from several problems. On the one hand, due to the constraint of a single view, resampling inevitably leads to high redundancy of the generated data. In particular, when only a few labeled objects are provided, the quality of the generated data is difficult to guarantee. On the other hand, the evaluation of unlabeled objects is judged by only the consistency of base classifiers, without considering their confidence and uncertainty. If base classifiers are weak, unlabeled objects may be mislabeled and classification noise is introduced. Therefore, it is highly desirable to improve the mechanism of training base classifiers and the strategy of selecting unlabeled objects.

3 Tri-trade for partially labeled data

In this section, the overall framework of the proposed model is first described. After that, a semi-supervised attribute reduction algorithm based on discernibility matrix is presented. Subsequently, the tri-trade model is proposed based on three distinct semi-supervised reducts. Finally, the effectiveness of the model is analyzed theoretically.

3.1 Overall framework

Tri-training is an efficient semi-supervised model that employs three classifiers to learn from unlabeled data. However, due to the single-view constraint, the tri-training model suffers from the problem of high redundancy of generated data after resampling. In fact, some datasets, particularly those with a large number of attributes, may generally be reduced to multiple attribute subsets, each of which can completely and competently represent the original data. In addition, these attribute subsets describe the original data from different perspectives, resulting in diverse induction biases. Therefore, by utilizing the diversity of multiple reduced subspaces, we can construct an effective multiview tri-trade model for partially labeled data, which is illustrated in Fig. 1.

Different from standard tri-training, the tri-trade model employs the attribute reduction technique to generate different views. More specifically, a semi-supervised discernibility matrix is first constructed for partially labeled data, and a heuristic algorithm is designed to generate three distinct semi-supervised reducts. On each reduct (view), a base classifier is trained using initially labeled data. Then, by utilizing data editing technique, two base classifiers select some confidently unlabeled data to update the third classifier. When no classifier can be updated, the algorithm terminates and yields a final classifier by combining the three refined classifiers. In the next sections, we elaborate on the details of the proposed model.

3.2 Discernibility matrix-based semi-supervised attribute reduction

In rough sets, traditional discernibility matrix-based attribute reduction methods are often used to deal with completely labeled or unlabeled data. However, in semi-supervised tasks, objects are only partially labeled. To deal with partially labeled data, a new discernibility matrix is developed. In a traditional discernibility matrix, discernible information is generated only from labeled objects or unlabeled objects. Intuitively, a reduct for partially labeled data should distinguish all kinds of objects. Thus, it is desirable that a semi-supervised discernibility matrix can consider both labeled and unlabeled objects. For this purpose, a semi-supervised discernibility matrix is constructed as follows.

Definition 9 Let $PS = (U = L \cup N, A = C \cup D, V', f)$ be a partially labeled data, the non-empty element of the semi-supervised discernibility matrix SM is defined as:

$$e_{ij} = \begin{cases} \left\{a \in C \mid a\left(x_i\right) \neq a\left(x_j\right)\right\}, & d\left(x_i\right) \neq d\left(x_j\right) \wedge \left(x_i \in L \wedge x_j \in L\right) \\ \left\{a \in C \mid a\left(x_i\right) \neq a\left(x_j\right)\right\}, & \left(x_i \in L \wedge x_j \in N\right) \vee \left(x_i \in N \wedge x_j \in L\right) \\ \left\{a \in C \mid a\left(x_i\right) \neq a\left(x_j\right)\right\}, & x_i \in N \wedge x_j \in N \end{cases} \tag{6}$$

In the definition, labeled objects with different labels are compared to generate discernible information. Due to the decision uncertainty of unlabeled data, all unlabeled objects are discerned from each other. In addition, to distinguish all kinds of objects, discernible information between labeled and unlabeled objects is generated. In the following, an example is given 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 1, where $U = \{x_1, x_2, \ldots, x_8\}$, $C = \{a_1, a_2, a_3, a_4, a_5\}$, $V_a = \{0, 1\}$ for every $a \in C$, and $V_D = \{d_1, d_2, ?\}$.

In Table 1, there are two labeled objects and six unlabeled objects. According to Definition 9, the discernibility matrix can be derived in Table 2. In Table 2, labeled objects with

Table 1 A partially labeled data

	a_1	a_2	a_3	a_4	a_5	d
x_1	1	1	1	0	0	d_1
x_2	0	0	0	1	1	d_2
x_3	1	1	1	1	1	?
x_4	0	0	0	0	0	?
x_5	1	1	0	0	0	?
x_6	0	0	1	1	1	?
x_7	1	0	0	0	0	?
x_8	1	1	1	1	0	?

different labels are compared to generate elements in $L \times L$; unlabeled objects are compared to generate elements in $N \times N$; labeled objects and unlabeled objects are compared to generate elements $L \times N$.

To perform semi-supervised attribute reduction based on the proposed discernibility matrix, we first introduce the concepts of the relevant set and the complement set of attributes:

Input: A partially labeled data $PS = (U = L \cup N, A = C \cup D, V', f)$
Output: A semi-supervised reduct set R
1: Compute the semi-supervised discernibility matrix SM of PS, $core \leftarrow \emptyset$, $R \leftarrow \emptyset$;
2: Add all singletons in SM to $core$ and initialize attribute subset queue $Queue \leftarrow core$;
3: **while** $Queue \neq \emptyset$ and $|R| \neq 3$ **do**
4: Get an attribute subset S from the head of $Queue$;
5: **if** $OM_{SM}(S) = \emptyset$ **then** Add S to R;
6: **else**
7: Add top three attributes with the minimum heuristic cost within $OM_{SM}(S)$ to S respectively, obtain S'_1, S'_2, S'_3;
8: Add S'_1, S'_2, S'_3 to $Queue$;
9: **end while**
10: **return** the set of semi-supervised reduct R.

Algorithm 1 Beam search algorithm for attribute reduction based on semi-supervised discernibility matrix.

Definition 10 Let $PS = (U = L \cup N, A = C \cup D, V', f)$ be a partially labeled data and SM be the semi-supervised discernibility matrix of PS. Then, for an attribute subset $B \subseteq C$, its relevant set is defined as:

$$RM_{SM}(B) = \bigcup \{e \in SM \mid \exists a \in B \wedge a \in e\} \tag{7}$$

Definition 11 Let $PS = (U = L \cup N, A = C \cup D, V', f)$ be a partially labeled data and SM be the semi-supervised discernibility matrix of PS. Then, for an attribute subset $B \subseteq C$, the complement set with respect to its relevant set is defined as:

$$OM_{SM}(B) = \{e - B | e \in RM_{SM}(B)\} \tag{8}$$

According to the above definitions, for an attribute set, its relevant set consistes of the elements that contain the attributes in the attribute set, and the relevant set after eliminating the attributes in the given attribute set comprises its complement set.

Based on the set operators presented above, an attribute reduction algorithm can be designed to generate reducts for partially labeled data. However, finding all reducts, or finding a minimal reduct, i.e., a reduct with the minimum number of attributes, is NP-hard. Thus, heuristic algorithms are preferred. By designing reasonable heuristic costs, heuristic algorithms can quickly obtain optimal reducts. Due to its simplicity and efficiency, the greedy forward search strategy of iteratively adding attributes is widely used in practical applications. However, greedy forward search tends to fall into local optimum. Beam search is a forward search heuristic algorithm utilized in this paper. It can explore several optimal reducts in parallel, which can be used as views for semi-supervised learning. More specifically, beam search utilizes a breadth-first strategy to find optimal reducts. In each iteration, candidate attributes are sorted according to the heuristic cost, and a certain number (called the beam width) of attributes with the minimum costs are preserved. Since the tri-trade model requires three distinct reducts to train three base classifiers, the beam width is set to three. In attribute reduction process, it is desired that attributes with strong discriminating power can be be preferentially selected, and optimal reducts should contain fewer attributes. Therefore, to evaluate the cost of each attribute, the heuristic function is defined as follows:

$$\text{heuristic cost}(a) = \frac{|S|}{|C|} + \frac{|RM_{SM}(a)|}{|SM|} \tag{9}$$

For an attribute a, the heuristic information consists of two parts: the ratio of the number of selected attributes $|S|$ in

the current attribute subset to the number of all attributes $|C|$, which aims at minimizing the number of attributes contained in the reducts, and the ratio of the number of elements in the attribute complement set $|RM_{SM}(a)|$ to the number of all elements $|SM|$, which aims at selecting the attributes with strong discriminating power. By using this heuristic information, the beam search algorithm can select the most important attributes to generate multiple reducts. This process can be described by Algorithm 1.

The algorithm starts with the construction of a semi-supervised discernibility matrix for partially labeled data. Since the core attributes have unique discriminating power (see Definition 7), the attribute subset is initialized with the core attributes to accelerate the search process (step 1 and step 2). In each iteration, the algorithm selects three attributes with the minimum heuristic costs and discards their relevant set. The search process terminates until the semi-supervised discernibility matrix is empty (step 3 to step 9). Finally, three optimal reducts are yielded, each of which has a nonempty intersection with any nonempty element of the semi-supervised discernibility matrix, thus maintaining the same discriminating power as all condition attributes.

Suppose that the partially labeled data contains $|U|$ objects described by $|C|$ condition attributes. The time cost of constructing a semi-supervised discernibility matrix is $O(|C||U|^2)$. In each iteration, the algorithm selects the optimal three attributes while deleting the corresponding relevant set from the matrix. In the worst case, after $|C|$ rounds of attribute selection, the matrix is empty. Therefore, the time cost of computing the optimal reducts based on the semi-supervised discernibility matrix is $O(|C|^2|U|^2)$. As a whole, Algorithm 1 has a total time cost of $O(|C||U|^2 + |C|^2|U|^2)$, which approximates $O(|C|^2|U|^2)$, and a total space cost of $O(|C||U|^2)$.

3.3 Multi-view tri-trade model for partially labeled data

In classic rough sets-based learning methods, the model typically employs a single classifier and mainly addresses labeled data. However, partially labeled data usually comprise relatively little labeled data and a considerable quantity of unlabeled data. When labeled data are limited, a learning model with a single classifier can hardly provide satisfactory results. Tri-training is a disagreement-based model that has been proven to be effective for partially labeled data [40]. Unfortunately, tri-training suffers from problems of the low diversity of base learners and poor quality of selected unlabeled data. Based on Algorithm 1, we can obtain three optimal reducts of partially labeled data. Since each reduct is a jointly sufficient attribute subset that can completely describe the overall data and the process of beam search that starts from different branches in parallel enables certain diversity among reducts, each reduct can be approximated as a sufficient and redundant view. Thus, we can utilize these reducts to improve tri-training.

In addition, not all unlabeled data are conducive to the learning model, and the selection of unlabeled objects is another key factor for the success of semi-supervised learning. Standard tri-training generates pseudo-labeled objects by majority voting. More specifically, if two classifiers make a consistent prediction on an unlabeled object, this object will be annotated with the pseudo-label and is considered a useful object to update the third classifier. However, in some circumstances (particularly in the early iteration), since initially labeled objects are insufficient to train strong base classifiers, a considerable number of objects may be wrongly classified.

The data editing technique is a commonly used method for error estimation, which aims to enhance the quality of training set by identifying and excluding mislabeled objects from the learning process. To improve the quality of training set, Zhang et al. [49] proposed a co-trade model based on data editing to improve co-training. The co-trade model first constructs a weighted graph over the labeled and unlabeled objects to describe the proximity in the attribute space using the k-nearest neighbor method. Based on the manifold assumption that objects with high similarity in the input space should have similar labels, the cut edge weight statistic is then used to explicitly evaluate the labeling confidence of unlabeled objects. Through data editing technique and co-training mechanism, the co-trade model can obtain high-quality labeled object sets to improve base classifiers. Motivated by the above fact, the data editing technique is introduced in tri-training to improve the quality of generating pseudo-labeled objects. More specifically, in each iteration of tri-training, two of the classifiers can use data editing technique to explicitly estimate the labeling confidence of unlabeled objects and collaboratively select unlabeled objects to generate pseudo-labels for the third classifier. This process is described by Algorithm 2.

In Algorithm 2, all parameters are initialized first (step 1 and step 2). In the iterative process, objects in unlabeled set are predicted first with classifiers clf_1 and clf_2 under each view (step 4). Then, a neighborhood graph is constructed to evaluate the labeling confidence explicitly (step 5). Under each view of unlabeled set, unlabeled data are sorted by labeling confidence in descending order, and an object subset N_i^* is chosen with the minimal expected prediction error ϵ_i'. Finally, two classifiers share labeling information to refine each other (step 6 and step 7). The iterative process terminates when the prediction error of either classifier increases on the original labeled set or when the expected prediction error of classifiers does not decrease. The last

Table 2 The semi-supervised discernibility matrix of partially labeled data in Table 1

		a_1	a_2	a_3	a_4	a_5			a_1	a_2	a_3	a_4	a_5
$L \times L$	e_{12}	1	1	1	1	1	$L \times N$	e_{13}	0	0	0	1	1
$N \times N$	e_{34}	1	1	1	1	1		e_{14}	1	1	1	0	0
	e_{35}	0	0	1	1	1		e_{15}	0	0	1	0	0
	e_{36}	1	1	0	0	0		e_{16}	1	1	0	1	1
	e_{37}	0	1	1	1	1		e_{17}	0	1	1	0	0
	e_{38}	0	0	0	0	1		e_{18}	0	0	0	1	0
	e_{45}	1	1	0	0	0		e_{23}	1	1	1	0	0
	e_{46}	0	0	1	1	1		e_{24}	0	0	0	1	1
	e_{47}	1	0	0	0	0		e_{25}	1	1	0	1	1
	e_{48}	1	1	1	1	0		e_{26}	0	0	1	0	0
	e_{56}	1	1	1	1	1		e_{27}	1	0	0	1	1
	e_{57}	0	1	0	0	0		e_{28}	1	1	1	0	1
	e_{58}	0	0	1	1	0							
	e_{67}	1	0	1	1	1							
	e_{68}	1	1	0	0	1							
	e_{78}	1	1	1	1	0							

two classifiers return pseudo-labeled objects with the same prediction on $N_1^* \cup N_2^*$ (step 13). Since the unlabeled sets N_1 and N_2 have been filtered by data editing technique and the initial classifiers clf_1 and clf_2 are improved after the co-training process, the final two classifiers can be combined to yield pseudo-labeled objects with high confidence.

Without loss of generality, assume that the partially labeled data has $|L|$ labeled objects and $|N|$ unlabeled objects described by $|C|$ attributes, and the time cost of training a base classifier is approximately $O(|C||U|)$. While the time cost of constructing the neighborhood graph is approximately $O(|U|^2)$ and the space cost is approximately $O(|U|^2)$. In each iteration, these two classifiers provide new pseudo-labeled objects for each other. Since the iterations can converge quickly, the time cost of training these classifiers is approximated as $O(|C||U|)$. Thus, the total time cost of Algorithm 2 is $O(|U|^2)$, and its space cost is $O(|U|^2)$.

To optimize the tri-training model, the tri-trade model is developed. By utilizing beam attribute reduction algorithm on a semi-supervised discernibility matrix, three distinct attribute subsets are generated from the original attribute set, on which three base classifiers are trained. By utilizing the proposed data editing technique, the quality of unlabeled objects is explicitly estimated and labeling information is reliably shared. The tri-trade model procedure is presented in Algorithm 3.

In Algorithm 3, three base classifiers are trained on three distinct reducts of the original condition attributes. After initializing all parameters, the classifiers iteratively learn from each other on unlabeled data. More specifically, in each round of tri-training, the classification error rate of each classifier is first estimated. Since it is difficult to estimate the classification error on unlabeled set, only the original labeled set is tested here, based on the heuristic assumption that the unlabeled objects have a similar distribution as the labeled objects. In detail, the estimated classification error rate is the proportion of the objects misclassified by both clf_j and clf_k to the objects consistently predicted by clf_j and clf_k. When there is no degradation in the performance of the combination of clf_j and clf_k , high confidence labeled objects are selected by data editing technology, and clf_i is updated by a certain number of newly labeled objects; otherwise, the classifier clf_i does not change. It should be noted that the data editing process has no impact on the three classifiers' updates. When no classifier can be updated, the algorithm terminates and yields the final classifier by combining the three retrained classifiers.

Assume the partially labeled data have $|L|$ labeled objects and $|N|$ unlabeled objects described by $|C|$ attributes where $|U| = |L| + |N|$. In each round of tri-training, two of the classifiers iteratively label objects for the third classifier using the data editing technique in Algorithm 2. The time cost of this process is $O(|U|^2 + |C||U|)$, and its space cost is $O(|C||U|)$. In the worst case, Algorithm 3 terminates after $|N|$ rounds of tri-training. Therefore, based on three distinct reducts of a given partially labeled data, the time cost of Algorithm 3 is at most $O(|U|^3)$ and its total space cost is approximated to $O(|C||U|)$.

Input: clf_1: the classifier under view I_1, clf_2: the classifier under view I_2; L_1: labeled set under view I_1, L_2: labeled set under view I_2; N_1: unlabeled set under view I_1, N_2: unlabeled set under view I_2; Maximum number of iterations $T = 30$.

Output: High-confidence objects in unlabeled set along with their predictive labels.

Set the expected prediction error of the two classifiers clf_1 and clf_2 on unlabeled set, $\epsilon_1 \leftarrow 1/\sqrt{L_1}$, $\epsilon_2 \leftarrow 1/\sqrt{L_2}$;
Test the error rates e_1 and e_2 of the two classifiers clf_1 and clf_2 on labeled set, $Iter \leftarrow 1$
while $Iter \leq T$ **do**
    Generate $clf_1(N_1)$ and $clf_1(N_2)$ // make predictions on unlabeled data;
    Construct neighbourhood graphs from $L_1 \cup clf_1(N_1)$ as well as $L_2 \cup clf_2(N_2)$ and estimate the labeling confidence based on the constructed graphs;
    Identify optimal choices $\left(N_1^*, \epsilon_1'\right)$ and $\left(N_2^*, \epsilon_2'\right)$ to generate pseudo-labels // clf_i has the lowest expected prediction error ϵ_i' on the subset N_i^* of the unlabeled set N_i;
    Retain clf_1' on $L_1 \cup clf_2(N_2^*)$ and clf_2' on $L_2 \cup clf_2(N_1^*)$;
    Test the error rates e_1' and e_2' of the two classifiers clf_1' and clf_2' on labeled set;
    if $(e_1' > e_1 \| e_2' > e_2)$ **or** $(\epsilon_1 > \epsilon_1' \& \epsilon_2' > \epsilon_2')$ **then** goto step 13;
    if $\epsilon_1' < \epsilon_1$ **then** $\epsilon_1 \leftarrow \epsilon_1', clf_1 \leftarrow clf_1'$;
    if $\epsilon_2' < \epsilon_2$ **then** $\epsilon_2 \leftarrow \epsilon_2', clf_2 \leftarrow clf_2'$;
end while
$clf_1' \leftarrow clf_1, clf_2' \leftarrow clf_2$;
return the pseudo-labeled objects with the same prediction of clf_1' and clf_2' on $N_1^* \cup N_2^*$.

Algorithm 2 The selection of high-confidence unlabeled data using data editing.

3.4 Theoretical analysis of the tri-trade model effectiveness

In the tri-training model, resampling data may deviate from the original data distribution. However, there is high redundancy in the generated data. Unlike the tri-training model, the tri-trade model trains base classifiers on three distinct reducts. From the perspective of attribute reduction, each reduct is a jointly sufficient subset of attributes and can preserve the same discriminating power as the original attribute set. Furthermore, the beam search attribute reduction algorithm ensures that the three reducts share as few attributes as possible; thus, each reduct describes the original data from a distinct view. The studies in [46] demonstrated that the co-training process can work well when the two classifiers have a large diversity, which guarantees the effectiveness of the proposed model for partially labeled data.

Another key factor for the success of tri-training is the quality of unlabeled objects. The tri-trade model employs the data editing technique to explicitly estimate the labeling confidence of unlabeled objects and utilizes the jointly predictive results of two classifiers. Additionally, a certain number of useful objects are selected for the third classifier to update only if the estimated performance of the classifier does not deteriorate. In essence, the principles of noise

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

Output: A combined classifier F

Generate three distinct semi-supervised reducts(views) I_1, I_2, and I_3 from the condition attribute set C by Algorithm 1;
Train three base classifiers clf_1, clf_2, and clf_3 under the views I_1, I_2, and I_3 using the labeled set, respectively;
Set the initial prediction error e_i and the number of newly labeled objects l_i of each classifier $e_i = 0.5$, $l_i = 0, i = (1, 2, 3)$;
repeat
    for $i = (1, 2, 3)$ **do**
        $S_i = \emptyset, update_i = False$;
        Test the error rate e_i' of the combination of clf_j and clf_k on L $//j, k \neq i$;
        if $e_i' < e_i$ **then**
            $L_i' = data\ editing(clf_j, clf_k, L_j, L_k, N_j, N_k)$ // use data editing technology in Algorithm 2 to generate a labeled set L_i' for clf_i;
            if $l_i = 0$ **then** $l_i = \left\lfloor \frac{e_i'}{e_i - e_i'} + 1 \right\rfloor$;
            if $l_i < |L_i'|$ **then**
                if $e_i'|L_i'| < e_i l_i$ **then** $update_i \leftarrow True$;
                else if $l_i > \frac{e_i'}{e_i - e_i'}$ **then**
                    $L_i' = subsample(L_i', \left\lfloor \frac{e_i'}{e_i - e_i'} + 1 \right\rfloor), update_i = True$ // to keep the inequality $e_i'|L_i'| < e_i l_i$;
            end if
        end if
    end for
    for $i = (1, 2, 3)$ **do**
        if $updat_i = True$ **then**
            Retrain clf_i on $L_i \cup L_i'$;
            $e_i \leftarrow e_i', l_i \leftarrow |L_i'|$;
    end for
until none of classifier changes;
return the combined classifier $F = combined(clf_1, clf_2, clf_3)$.

Algorithm 3 Tri-trade model for partially labeled data.

learning are implicitly embedded in the tri-training model. According to noise learning theory [49], the following formula holds:

$$m = \frac{c}{\epsilon^2(1-2\eta)^2}, \tag{10}$$

where ϵ is the expected worst classification error rate, η denotes the upper bound of the classification noise rate, and c is a constant for a specific learning task. By reformulating inequality (10), the following utility function for the classification noise rate can be derived:

$$u = \frac{c}{(1-2\eta)^2} = m\epsilon^2, \tag{11}$$

To lower the classification noise rate, the utility function should be reduced in each iteration, i.e., $u' < u$. The following inequality can be obtained:

$$m'\epsilon'^2 < m\epsilon^2, \tag{12}$$

Since the iteration process satisfies $\epsilon' < \epsilon$, inequality (12) can be converted as follows:

$$m'\epsilon' < m\epsilon, \tag{13}$$

and the following constraints can be derived:

$$0 < \frac{\epsilon'}{\epsilon} < \frac{m}{m'} \tag{14}$$

Note that $m'\epsilon'$ may not be smaller than $m\epsilon$ since m' may be much larger than m. In this case, the function $subsample(L_i', \left\lceil \frac{\epsilon l_i}{\epsilon_i'} - 1 \right\rceil)$ randomly selects a certain number of objects L_i'. Let the integer s denote the size of L_i' after subsampling. If it satisfies:

$$s = \left\lceil \frac{m\epsilon}{\epsilon'} - 1 \right\rceil \tag{15}$$

the constrained condition of inequality (13) can be satisfied as well. In this case, m needs to satisfy the following condition:

$$m > \frac{\epsilon'}{\epsilon' - \epsilon}, \tag{16}$$

According to inequality (14), the proposed tri-trade model considers the classifier to be updated on some unlabeled objects only when the estimated error rate does not increase. According to inequalities (15) and (16), the classifier selects a certain number of unlabeled objects in each iteration to satisfy the constraint of inequality (13), thus reducing (or at least maintaining) the classification noise rate. Therefore, the tri-trade model can make efficient use of unlabeled data to enhance its performance.

4 Empirical analysis

This experiment serves two purposes. One is to evaluate the effectiveness of the discernibility matrix-based semi-supervised attribute reduction algorithm. The other aims to compare the performance of the proposed model with other semi-supervised methods. All experiments were carried out on a Windows 10 machine with an Intel(R) Core (TM) i7-10700 CPU @ 2.90GHz dual-core processor and 32GB RAM, and all codes were implemented by Python 3.7 in the platform of PyCharm.

4.1 Investigated datasets and experimental design

In the experiment, twelve UCI datasets are tested. It should be noted that some of the datasets are multicategory. To construct binary classification datasets, the category with the most objects is treated as one class, which is referred to as the positive class, and the remaining objects are grouped into another class, which is referred to as the negative class. Table 3 reports detailed information about all datasets. The first column of the table contains the name of the selected datasets, the second column $|C|$ and the third column $|U|$ are the number of attributes and objects in each dataset, respectively, the fourth column "POS/NEG" gives the percentage of positive objects and negative objects, the fifth column "Missing" indicates whether the dataset has missing values, and the last column "Inconsistency" records the number of inconsistent objects in these datasets. Note that these datasets do not have multiple and redundant views.

To facilitate the experiments, the missing values of attributes in each dataset are replaced by the average or most frequent values of the respective attributes. Since the proposed model is designed for partially labeled data with categorical attributes, each numerical attribute must be discretized into categorical attributes. Due to the simplicity and effectiveness, equal frequency binning with the three bins [50] is employed in the experiments.

To evaluate the performance of the proposed method, ten 10-fold cross-validation tests are employed in the experiments. In each fold, 90% of the objects are selected as the training set, while the remaining 10% are treated as the test set. According to the label rates, the training set is further randomly divided into a set of labeled objects L and a set of unlabeled objects N. The label rates include 1%, 5%, 10%, 15%, and 20%. Under each label rate, the training set is divided ten times independently and randomly. For instance, assuming that there are 1000 objects, 900 objects in each fold are selected as the training set and the remaining 100 objects are regarded as the test sets. When the label rate is 10%, 90 objects with labels are placed in the labeled set

Table 3 Investigated datasets

Datasets	$\|C\|$	$\|U\|$	POS/NEG	Missing	Inconsistency
Anneal(anneal)	38(6)	898	76.17%-23.83%	N	11
Biodegradation(biodegradation)	41(41)	1055	66.26%-33.74%	N	15
Colic(colic)	22(7)	368	63.04%-36.96%	Y	0
Credit-a(credit-a)	15(6)	690	55.51%-44.49%	Y	6
Credit-g(credit-g)	20(7)	1000	70.00%-30.00%	N	2
Gesture-phase-a1va3(gesture1)	32(32)	1743	60.18%-39.82%	N	33
Gesture-phase-a2va3(gesture2)	32(32)	1260	61.19%-38.81%	N	27
Parkinson-speech-train(parkinson)	26(26)	1040	50.00%-50.00%	N	79
Polish-companies-bankruptcy-1year(polish)	64(64)	7027	96.14%-3.86%	Y	2
Spambase(spambase)	57(57)	4601	60.60%-39.40%	N	141
Turkiye-student-evaluation-specific(turkiye)	31(31)	5820	61.87%-38.13%	N	3349
Wall-following navigation task(wall)	24(24)	5456	59.59%-40.41%	N	306

L, while the remaining 810 objects after removing labels are placed in the unlabeled set N, and the data division between L and N is repeated ten times independently and randomly. Here, the ratio POS/NEG in the L, N and test set are maintained to be consistent with the original dataset.

To investigate the effectiveness of the beam attribute reduction algorithm for partially labeled data, all selected datasets are tested at a label rate of 10%, and the results of attribute reduction in ten times 10-fold cross-validation are collected. Table 4 shows the statistical results, including the maximum, minimum, and average number of attributes in the reducts, which are listed in the third, fourth, and fifth columns, respectively. In addition, the real reduct information, i.e., attribute reduction at 100% label rate, is collected for comparison. The last column provides a comparison between the semi-supervised reduct and the ground-truth supervised reduct, denoted as the approximate rate, i.e., the rate of the average number of attributes in the semi-supervised reduct to that in the ground-truth supervised reduct.

Table 4 shows that the number of attributes is reduced after semi-supervised attribute reduction. It is worth mentioning that core attributes in the reducts are always retained. The reason may be that in the semi-supervised discernibility matrix, some objects must be discriminated by core attributes. Compared to the ground-truth reduct, the average approximation rate of semi-supervised attribute approximation across all datasets is 72.87%. Notably, at a label rate of 10%, the approximation rates on datasets "credit-a", "parkinson", "turkiye" and "wall" are greater than 80%. These results indicate the effectiveness of the

Table 4 Results of semi-supervised attribute reduction under the label rate of 10%

Dataset	Raw	Semi-supervised reduct			Ground-truth reduct			Approximate rate
		Max	Min	Average	Max	Min	Average	
Anneal	38	24	23	23.7	14	13	14.7	62.03%
Biodegradation	41	25	23	23.9	15	14	14.8	61.92%
Colic	22	14	12	12.9	9	8	8.8	68.22%
Credit-a	15	13	13	13.0	11	10	10.8	83.08%
Credit-g	20	14	12	13.1	10	9	9.9	75.57%
Gesture1	32	28	26	26.7	19	17	18.5	69.29%
Gesture2	32	26	23	24.3	17	16	16.6	68.31%
Parkinson	26	20	19	19.4	16	14	15.7	80.93%
Polish	64	33	30	31.2	16	14	15.1	48.40%
Spambase	57	55	52	54.2	38	35	36.8	67.90%
Turkiye	31	31	29	30.2	30	29	29.2	96.69%
Wall	24	22	21	21.6	20	19	19.9	92.13%
Avg.	33.5	25.4	23.6	24.5	17.9	16.5	17.6	72.87%

proposed attribute reduction method for partially labeled data.

4.2 The effectiveness of the tri-trade method

To evaluate the performance of the proposed tri-trade model, we compare it with the supervised Laplace score and unsupervised Fisher score. Both are standard filter-style methods that assign a score to each attribute and then select the k attributes with the highest score. The main idea of Fisher score [51] is to identify attributes with strong distinguishing power, which is reflected as small intraclass distance and large interclass distance. The main idea of Laplace score [52] is to construct the nearest neighbor graph over all data, and the importance of each attribute is evaluated according to its locality-preserving power [50]. In addition, we compare the proposed tri-trade model to traditional semi-supervised methods such as self-training, co-training, co-trade, and tri-training. Self-training [48] is a self-taught method with only a single classifier. The initial classifier is trained on labeled data and is iteratively refined by its most confident self-labeled data. Co-training [23] is a multi-view disagreement-based method. It trains two initial classifiers on two attribute sets, and one classifier updates the other one with high-confidence objects in each iteration. Since most datasets lack naturally partitioned views, the requirements of the co-training are difficult to satisfy. However, it has been proven that [46] unlabeled objects can still improve the performance of co-training by randomly splitting the original attribute set into two subsets. Therefore, in the experiments, the attributes in each dataset are randomly partitioned into two disjoint sets of nearly equal size. Moreover, co-trade [49] is improved on co-training. It selects high-quality unlabeled objects by a data editing technique to refine the base classifiers. Tri-training [40] is also a multiview disagreement-based method but uses three base classifiers. The settings for all selected methods are shown in Table 5.

In Table 5, to demonstrate the potential of the proposed tri-trade model, full supervised learning, i.e., under the label rate of 100%, is set up for comparison. In addition, supervised learning is also performed on the reduct obtained from attribute reduction based on the Laplace score over $L \cup N$ and Fisher score over L, respectively. The number of attributes k remains the same as the optimal reduct of the semi-supervised discernibility matrix-based method. In addition, self-training trains its base classifier on the optimal reduct obtained from attribute reduction based on semi-supervised discernibility matrix over $L \cup N$. In co-training and co-trade, two views are generated by randomly dividing the original attribute set into two disjoint subsets with equal size. Tri-training obtains three labeled object subsets by resampling over L, while the tri-trade model obtains three views by semi-supervised discernibility matrix-based attribute reduction over $L \cup N$. In the proposed tri-trade model, the maximum number of iterations of data editing is set to 30. However, empirical results reveal that the training process terminates no more than 10 rounds in most cases. It should be noted that, compared to the co-training, co-trade, and tri-training, the tri-trade model uses a subset of attributes rather than all of them.

In the experiments, the label rate is set to 10%. Two different base classifiers, J48 and Naive Bayes, are employed and ten 10-fold cross-validations are performed to evaluate the performance. The average classification error rates of the selected methods are recorded in Tables 6 and 7. The column "Max" indicates the average error rates of full supervised learning on different datasets, and the third to ninth columns represent the average error rates of other methods in Table 5. The "Avg." row shows the average classification error rate of each method on selected datasets. The best classification results among all methods on each dataset are highlighted in bold.

As shown in Tables 6 and 7, there is a significant difference in the performance among the selected methods. By comparing the results, when evaluated by the number of datasets with the best classification performance, the proposed tri-trade model is always the winner. More specifically, when using J48, the tri-trade model wins 7 out of 12 datasets, while other methods win at most 2

Table 5 Experimental settings

Setting	Method
Full supervised learning	Ground truth
Choose top k attributes over $L \cup N$	Laplace score
Choose top k attributes over L	Fisher score
Discernibility matrix-based attribute reduction over $L \cup N$(1 view)	Self-training
Randomly divide all attributes with equal size (2 views)	Co-training
Randomly divide all attributes with equal size (2 views)	Co-trade
Resampling over L	Tri-training
Discernibility matrix-based attribute reduction over $L \cup N$(3 views)	Tri-trade

Table 6 Average performance of the selected methods using J48 classifier at 10% label rate

Datasets	Max	Laplace score	Fisher score	Self-training	Co-training	Co-trade	Tri-training	Tri-trade
Anneal	0.0300	0.1394	0.2246	0.1070	0.1473	0.1261	0.0791	**0.0776**
Biodegradation	0.1974	0.2540	0.2596	0.2695	0.2294	0.2800	**0.2212**	0.2414
Colic	0.2559	0.3203	0.2717	0.3044	0.3728	0.2997	**0.2669**	0.2872
Credit-a	0.2029	0.2285	0.2531	0.2207	0.2810	0.2478	0.2079	**0.2019**
Credit-g	0.3280	0.3660	0.3733	0.3760	0.4312	**0.3062**	0.3401	0.3368
Gesture1	0.2697	0.2514	0.2543	0.2643	0.2450	0.2460	0.2304	**0.2008**
Gesture2	0.3230	0.2886	0.2850	0.2969	0.2648	0.2734	0.2738	**0.2318**
Parkinson	0.4490	0.4424	0.4426	0.4494	0.4411	0.4409	0.4388	**0.4324**
Polish	0.0470	0.0718	0.0688	0.0662	0.0699	0.1100	0.0603	**0.0571**
Spambase	0.1765	**0.1251**	0.1340	0.1460	0.1340	0.1636	0.1276	0.1590
Turkiye	0.3789	0.4079	0.4096	0.4132	0.4310	**0.3907**	0.4141	0.3933
Wall	0.1310	0.1454	0.1448	0.1550	0.1861	0.1908	0.1546	**0.1432**
Avg	0.2324	0.2534	0.2601	0.2557	0.2680	0.2563	0.2346	**0.2302**

datasets; when using Naive Bayes, the tri-trade model wins 8 out of 12 datasets, while other methods win at most 2 datasets. When evaluated by the average classification error rate, the tri-trade model has an average classification error rate of 23.02% when using J48, which outperforms the full supervised method (23.24%). In contrast, all other methods perform worse than the tri-trade model. Impressively, the average classification error rate of the tri-trade model is 27.35% when using Naive Bayes, which is even better than fully supervised method (29.66%). In summary, the classification performance of the tri-trade model outperforms other semi-supervised methods and is even better than that of the fully supervised method. These results indicate that the tri-trade model can effectively exploit unlabeled data to enhance its performance.

To further evaluate the potential of the proposed model, the methods in Table 5 are performed at other label rates, including 1%, 5%, 10%, 15%, and 20%. Figures 2 and 3 show the average error rates of all methods. Note that "Max" refers to the performance of a single classifier with a label rate of 100%.

As shown in the figures, both the Fisher score and Laplace score perform poorly on most datasets, since their reducts have a loss of discernibility and they do not utilize unlabeled data to improve the classifier. For instance, in Figs. 2(a), 3(b), (c), and (e), the performance of both Fisher score and Laplace score is far inferior to other methods. Self-training is a single view model and unlabeled data are self-labeled. In general, self-training yields fewer desirable outcomes, as illustrated in Figs. 2(g), (h), 3(f), (g), (h) and (l). One reason may be that the

Table 7 Average performance of the selected methods using Naive Bayes classifier at 10% label rate

Datasets	Max	Laplace score	Fisher score	Self-training	Co-training	Co-trade	Tri-training	Tri-trade
Anneal	0.1581	0.3664	0.4843	0.3749	0.3318	**0.3088**	0.3421	0.4281
Biodegradation	0.3814	0.4254	0.4059	0.3010	0.3775	0.3943	0.3817	**0.2861**
Colic	0.3269	0.3250	0.2989	**0.2686**	0.3256	0.3047	0.3261	0.2768
Credit-a	0.1856	0.1925	0.2431	0.1818	0.1871	0.1906	0.1860	**0.1743**
Credit-g	0.5381	0.4755	0.5517	0.3588	0.5302	0.4560	0.5378	**0.3104**
Gesture1	0.1728	0.1782	0.1771	0.1914	0.1732	0.1739	**0.1723**	0.1744
Gesture2	0.2114	0.2135	0.2151	0.2259	**0.2104**	0.2109	0.2110	**0.2104**
Parkinson	0.4231	0.4219	0.4274	0.4862	0.4188	0.4188	0.4282	**0.4135**
Polish	0.3340	0.2697	0.2598	**0.1987**	0.2712	0.2731	0.3156	0.2038
Spambase	0.1735	0.2300	0.2269	0.2458	0.2228	0.2337	0.2243	**0.1778**
Turkiye	0.4172	0.4177	0.4179	0.4174	0.4132	0.4112	0.4179	**0.4046**
Wall	0.2370	0.2373	0.2374	0.3406	0.2378	0.2378	0.2379	**0.2218**
Avg	0.2966	0.3128	0.3288	0.2993	0.3088	0.3011	0.3152	**0.2735**

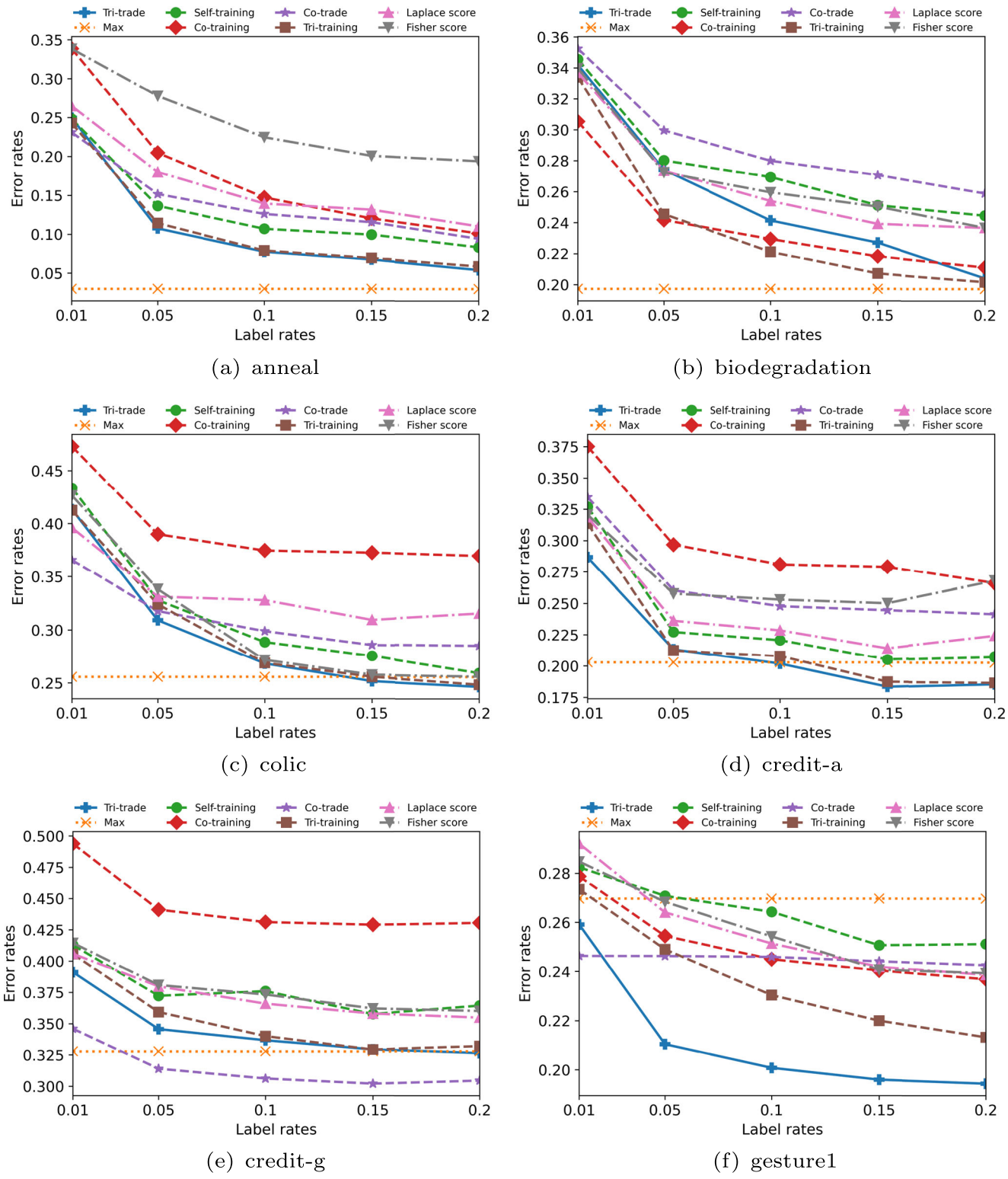
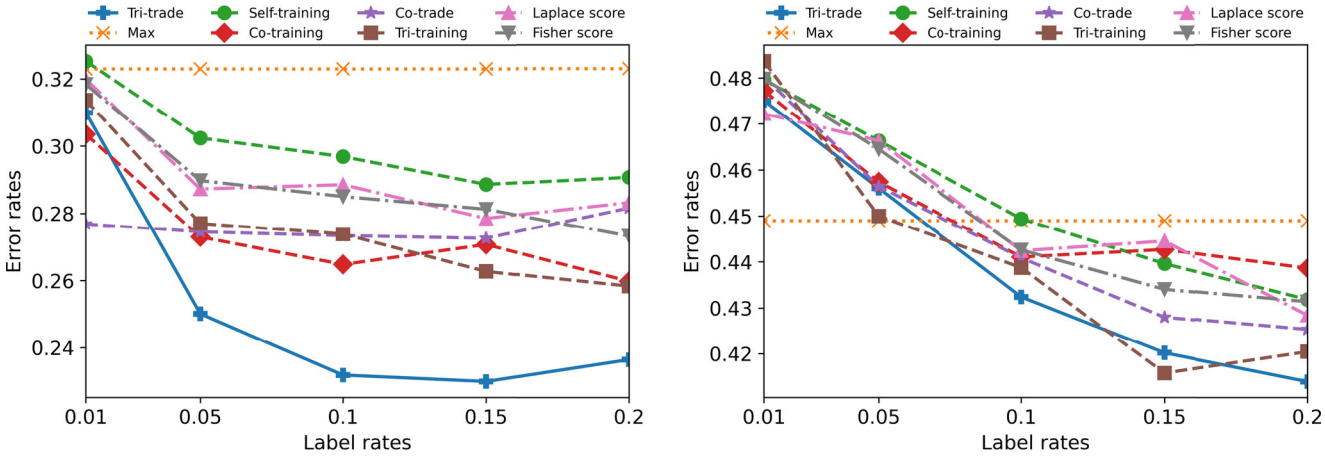
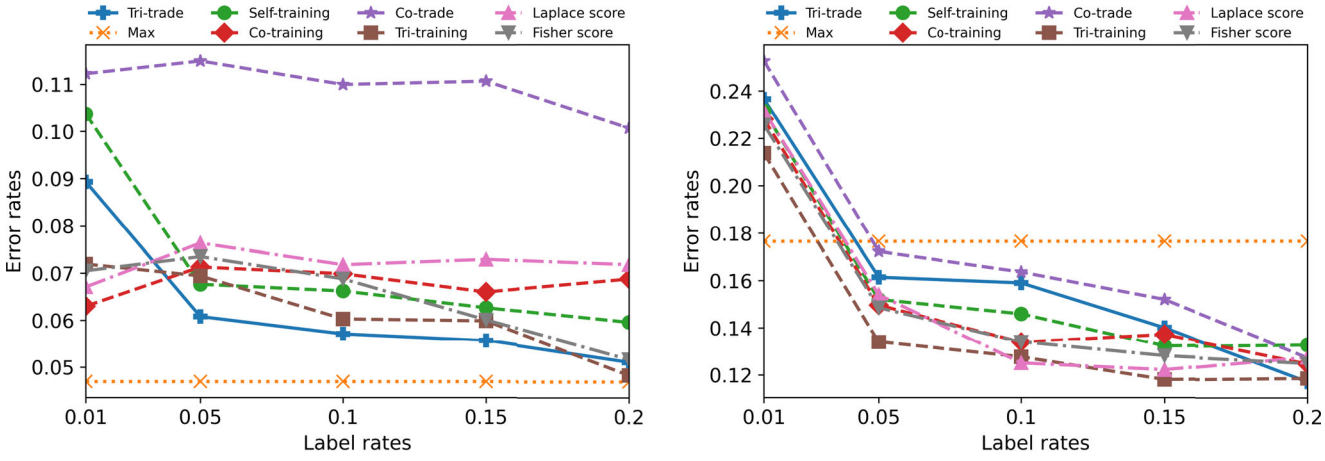


Fig. 2 Average error rates of the selected methods under different label rates (J48)

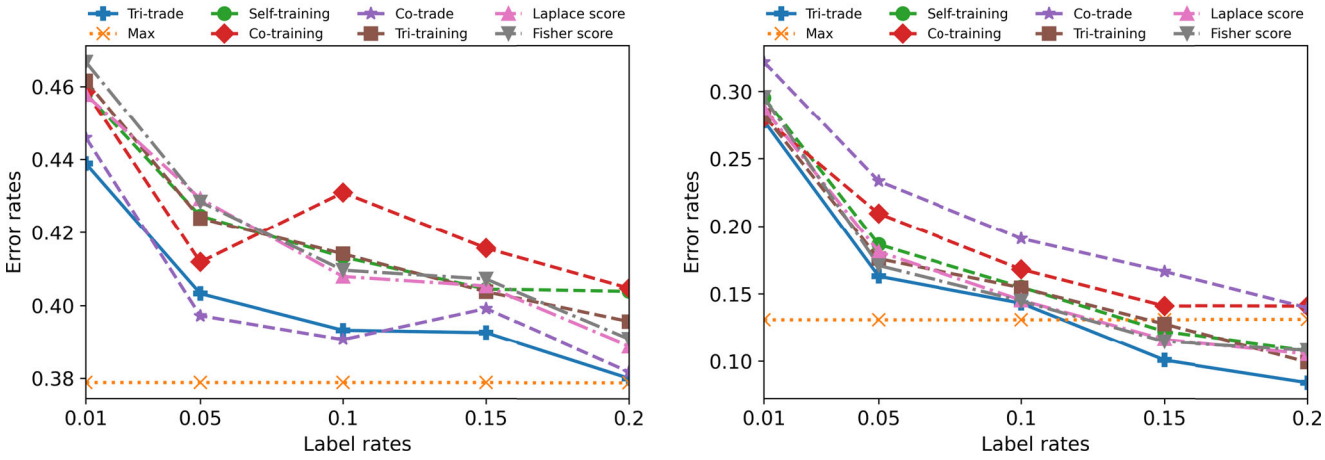
initially labeled data are not representative, and thus the generalization ability of the base classifier is unstable. Furthermore, the utilization of unlabeled objects affects performance. Self-training inevitably mislabeled objects, which further degrades performance. Co-training and co-trade are disagreement-based learning methods that employ two classifiers. However, their overall performance is not satisfactory. The main reason is that the co-training



(g) gesture2 (h) parkinson



(i) polish (j) spambase



(k) turkiye (l) wall

Fig. 2 (continued)

paradigm requires the original attribute set to be naturally partitioned, while in the experiments the subspaces for these two classifiers are formed by randomly dividing the whole attribute set by half. Obviously, this does not guarantee the quality of these two base classifiers. Therefore, the label information exchanged by these two

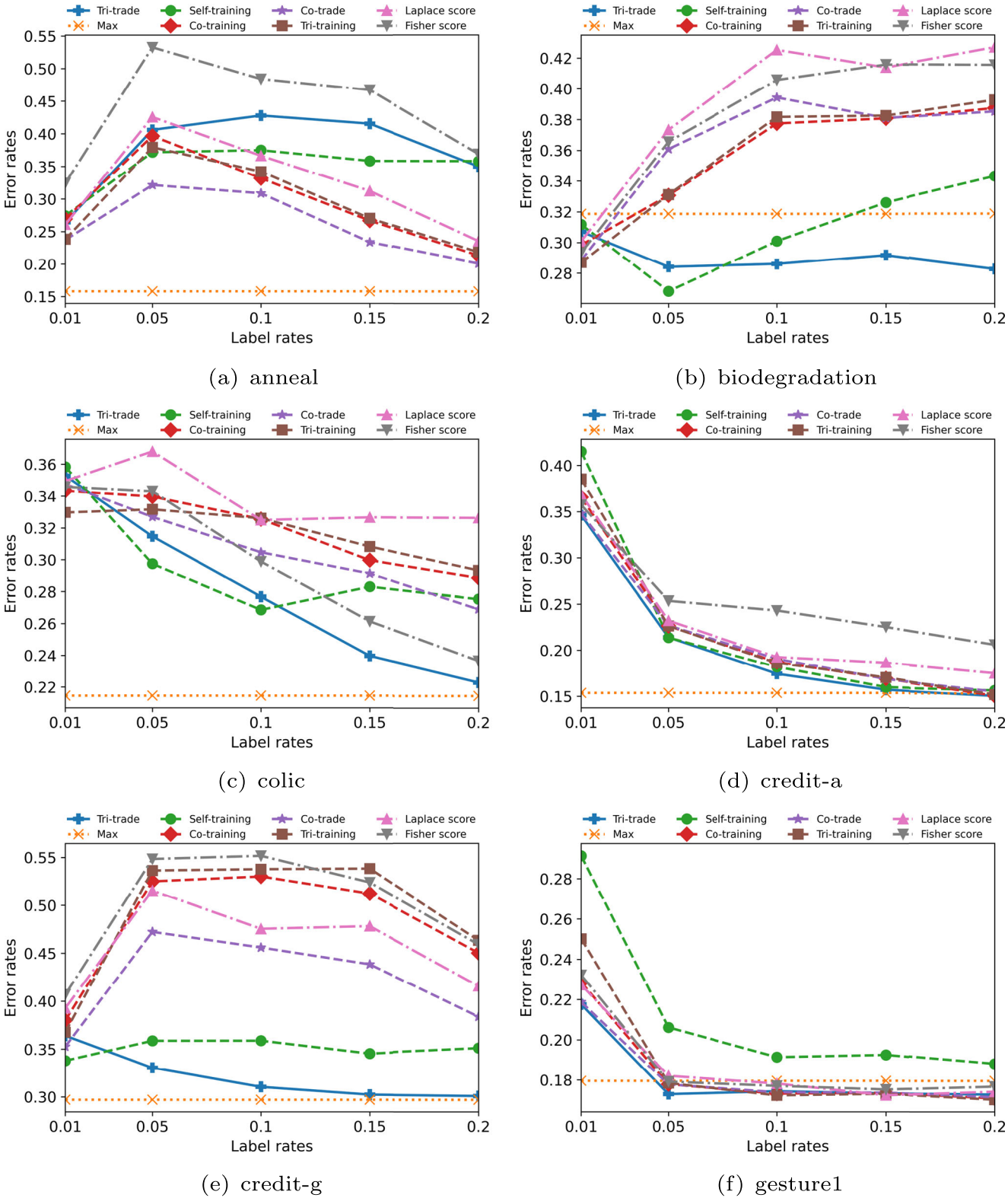
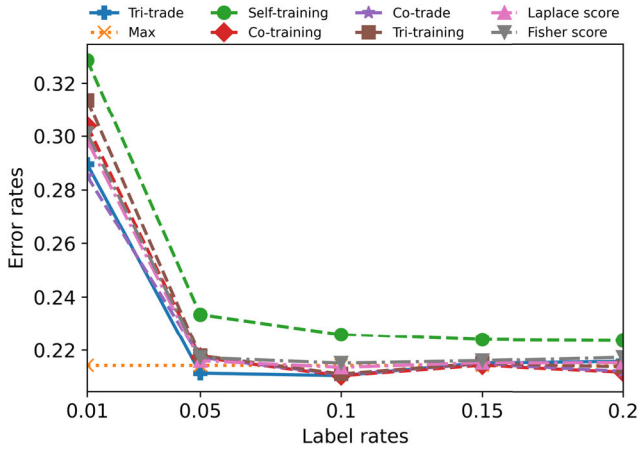
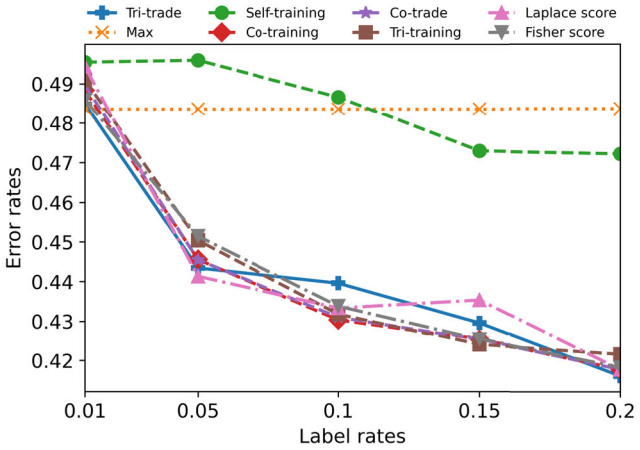


Fig. 3 Average error rates of the selected methods under different label rates (Naive Bayes)

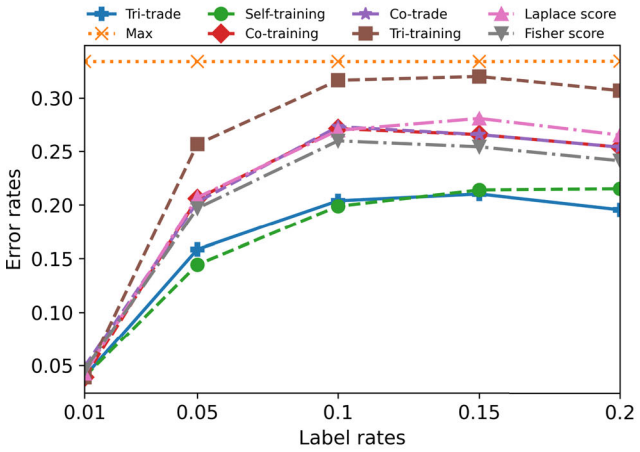
classifiers may contain noise, even though the co-trade imposes a restriction on the exchange of unlabeled objects, resulting in poor performance. Figure 2(c), (d) and (l) demonstrate this trend. Tri-training employs three classifiers to determine how to select unlabeled objects for labeling. The performance of the method remains deficient. The



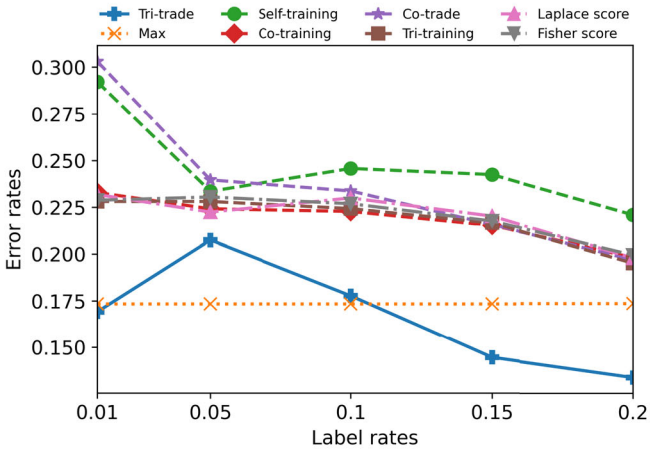
(g) gesture2



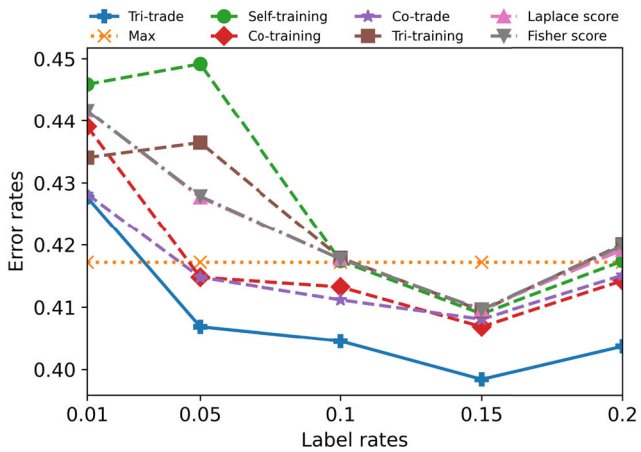
(h) parkinson



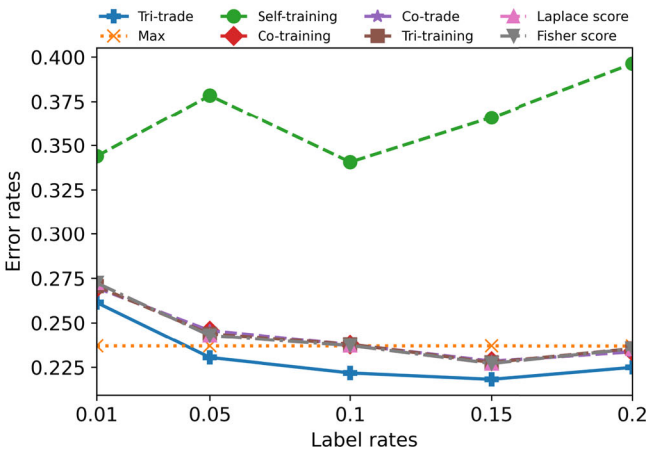
(i) polish



(j) spambase



(k) turkiye



(l) wall

Fig. 3 (continued)

reason may be twofold. The quality of resampling data is not guaranteed, leading to the lack of diversity in the generated classifiers. Additionally, the pseudo-labels produced by majority voting are insufficiently accurate. These lead to unstable performance of tri-training, which can be observed in Fig. 3(e), (i) and (k). Compared to the resampling operation in tri-training, tri-trade model trains its base classifiers on distinct reduced subspaces, each of which is a sufficient attribute subset that maintains the same discriminating power as the whole attribute set. By exploiting unlabeled data, tri-trade model achieves impressive performance. Tri-trade model estimates the labeling confidence explicitly and generates the pseudo-labeled objects by two enhanced classifiers. The tri-trade model carefully selects unlabeled objects for learning, and base classifiers are updated only when the pseudo-labeled labeled objects have a positive influence. Therefore, the tri-trade model can enhance performance by utilizing truly useful unlabeled objects.

Overall, the proposed tri-trade model outperforms all other methods under different label rates. Note that on some large datasets, such as "polish", "turkiye" and "wall", the number of labeled objects is sufficient to train a powerful classifier when the label rate is below 20%. However, even in these cases, tri-trade model is still effective. In addition, the performance on some datasets is especially excellent, as seen in Figs. 2(f), (g) and 3(j), where the average error rate of tri-trade model is much lower than that of the other methods. These experimental results demonstrate the superiority of semi-supervised attribute reduction as well as the data editing technique, showing that the tri-trade model has great potential to learn from partially labeled data.

5 Conclusion

In many real-world scenarios, unlabeled data are massive while labeled data are scarce. The strategy of selecting and utilizing unlabeled data is essential for the learning model of partially labeled data. In this study, a novel tri-trade model is proposed for partially labeled data. To obtain multiple distinct views from partially labeled data, a semi-supervised attribute reduction algorithm based on discernibility matrix is developed. Moreover, a new data editing technique is introduced to explicitly estimate the labeling confidence and to cautiously select unlabeled objects to improve the base classifiers. Theoretical analysis and comparative experiments on UCI datasets reveal that the proposed tri-trade model has prominent performance when compared to other methods. Admittedly, the proposed model is only applicable to partially labeled data with categorical attributes, which means that the numerical attributes must be discretized. Extending the model to deal with partially labeled data with both categorical and numerical attributes is worth investigating in the future. Additionally, it is worthwhile to explore other effective strategies for evaluating the labeling confidence of unlabeled data.

Acknowledgements The authors would like to thank the Editor-in-Chief, Editor, and anonymous reviewers for their kind help and valuable comments. This work is supported in part by the National Natural Science Foundation of China (Nos. 61806127, 62076164), the Natural Science Foundation of Guangdong Province, China (No. 2021A1515011861), Shenzhen Science and Technology Program (No. JCYJ20210324094601005), and Shenzhen Institute of Artificial Intelligence and Robotics for Society.

Data Availability Data and code are available from the corresponding author upon reasonable request.

Declarations

Competing interests The authors declare that they have no competing interests that could influence the work reported in this paper.

References

1. Pawlak Z (1982) Rough sets. Int J Comput Inf Sci 11(5):341–356
2. Xu W, Yu J (2017) A novel approach to information fusion in multi-source datasets: a granular computing viewpoint. Inf Sci 378:410–423
3. Chen X, Xu W (2022) Double-quantitative multigranulation rough fuzzy set based on logical operations in multi-source decision systems. Int J Mach Learn Cybern 13(4):1021–1048
4. Xue Z, Zhang R, Qin C, Zeng X (2018) A rough ν-twin support vector regression machine. Appl Intell 48(11):4023–4046
5. Sun L, Zhang X, Qian Y, Xu J, Zhang S, Tian Y (2019) Joint neighborhood entropy-based gene selection method with fisher score for tumor classification. Appl Intell 49(4):1245–1259
6. Pawlak Z (1991) Rough sets: theoretical aspects of reasoning about data. Kluwer Academic Publisher, Dordrecht
7. Bai S, Lin Y, Lv Y, Chen J, Wang C (2021) Kernelized fuzzy rough sets based online streaming feature selection for large-scale hierarchical classification. Appl Intell 51(3):1602–1615
8. Li Y, Cai M, Zhou J, Li Q (2022) Accelerated multi-granularity reduction based on neighborhood rough sets. Appl Intell 52(15):17636–17651
9. Sun L, Zhang J, Ding W, Xu J (2022) Mixed measure-based feature selection using the fisher score and neighborhood rough sets. Appl Intell 52:17264–17288
10. Wang Cz, Huang Y, Ding W, Cao Z (2021) Attribute reduction with fuzzy rough self-information measures. Inf Sci 549:68–86
11. Wang C, Qian Y, Ding W, Fan X (2022) Feature selection with fuzzy-rough minimum classification error criterion. IEEE Trans Fuzzy Syst 30(8):2930–2942
12. Wang C, Huang Y, Shao M, Hu Q, Chen D (2020) Feature selection based on neighborhood self-information. IEEE Trans Fuzzy Syst 50(9):4031–4042
13. Zhang X, Yao Y (2022) Tri-level attribute reduction in rough set theory. Expert Syst Appl 190:116–187
14. Zhang X, Jiang J (2022) Measurement, modeling, reduction of decision-theoretic multigranulation fuzzy rough sets based on three-way decisions. Inf Sci 607:1550–1582

15. Yang X, Li M, Fujita H, Liu D, Li T (2022) Incremental rough reduction with stable attribute group. Inf Sci 589:283–299
16. Liu K, Li T, Yang X, Ju H, Yang X, Liu D (2022) Hierarchical neighborhood entropy based multi-granularity attribute reduction with application to gene prioritization. Int J Approx Reason 148:57–67
17. Cai M, Lang G, Fujita H, Li Z, Yang T (2019) Incremental approaches to updating reducts under dynamic covering granularity. Knowl-Based Syst 172:130–140
18. Yang X, Yang Y, Luo J, Liu D, Li T (2022) A unified incremental updating framework of attribute reduction for two-dimensionally time-evolving data. Inf Sci 601:287–305
19. Yang X, Li Y, Liu D, Li T (2022) Hierarchical fuzzy rough approximations with three-way multi-granularity learning. IEEE Trans Fuzzy Syst 30(9):3486–3500
20. Wei W, Wu X, Liang J, Cui J, Sun Y (2018) Discernibility matrix based incremental attribute reduction for dynamic data. Knowl-Based Syst 140:142–157
21. Ma F, Ding M, Zhang T, Cao J (2019) Compressed binary discernibility matrix based incremental attribute reduction algorithm for group dynamic data. Neurocomputing 344:20–27
22. Liu Y, Zheng L, Xiu Y, Yin H, Zhao S, Wang X, Chen H, Li C (2020) Discernibility matrix based incremental feature selection on fused decision tables. Int J Approx Reason 118:1–26
23. Gao C, Zhou J, Miao D, Wen J, Yue X (2021) Three-way decision with co-training for partially labeled data. Inf Sci 544:500–518
24. Xin X, Shi C, Sun J, Xue Z, Song J, Peng W (2022) A novel attribute reduction method based on intuitionistic fuzzy three-way cognitive clustering. Appl Intell :1–15
25. Wu F, Jing X, Wei P, Lan C, Ji Y, Jiang G, Huang Q (2022) Semi-supervised multi-view graph convolutional networks with application to webpage classification. Inf Sci 591:142–154
26. Idhammad M, Afdel K, Belouch M (2018) Semi-supervised machine learning approach for ddos detection. Appl Intell 48(10):3193–3208
27. Mittal H, Pandey AC, Pal R, Tripathi A (2021) A new clustering method for the diagnosis of covid19 using medical images. Appl Intell 51(5):2988–3011
28. Dai J, Hu Q, Zhang J, Hu H, Zheng N (2016) Attribute selection for partially labeled categorical data by rough set approach. IEEE Trans Cybern 47(9):2460–2471
29. Hu S, Miao D, Zhang Z, Luo S, Zhang Y, Hu G (2018) A test cost sensitive heuristic attribute reduction algorithm for partially labeled data. In: International joint conference on rough sets, Springer, pp 257–269
30. Xie X, Qin X, Huang G, Zhao W (2019) Attribute reduction for partially labeled data based on hypergraph models. In: 2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI), pp 1434–1439
31. Liu K, Yang X, Yu H, Mi J, Wang P, Chen X (2019) Rough set based semi-supervised feature selection via ensemble selector. Knowl-Based Syst 165:282–296
32. Gao C, Zhou J, Miao D, Yue X, Wan J (2021) Granular-conditional-entropy-based attribute reduction for partially labeled data with proxy labels. Inf Sci 580:111–128
33. Wang R, Chen D, Kwong S (2013) Fuzzy-rough-set-based active learning. IEEE Trans Fuzzy Syst 22(6):1699–1704
34. Min F, Liu F-L, Wen L-Y, Zhang Z-H (2019) Tri-partition cost-sensitive active learning through knn. Soft Comput 23(5):1557–1572
35. Cekik R, Uysal AK (2020) A novel filter feature selection method using rough set for short text data. Expert Syst Appl 160:113691
36. Kuo C, Shieh H (2015) A semi-supervised learning algorithm for data classification. Int J Pattern Recogn Artif Intell 29(05):1551007
37. Bharadwaj A, Ramanna S (2019) Categorizing relational facts from the web with fuzzy rough sets. Knowl Inf Syst 61(3):1695–1713
38. Agrawal S, Ahmed R, Anand Kumar M, Ramanna S (2022) Categorizing relations via semi-supervised learning using a hybrid tolerance rough sets and genetic algorithm approach. In: Soft computing for data analytics, classification model, and control, Springer, pp 103–116
39. Bougoudis I, Demertzis K, Iliadis L, Anezakis V-D, Papaleonidas A (2018) Fussffra, a fuzzy semi-supervised forecasting framework: the case of the air pollution in athens. Neural Comput Applic 29(7):375–388
40. Zhou Z, Li M (2005) Tri-training: exploiting unlabeled data using three classifiers. IEEE Trans Knowl Data Eng 17(11):1529–1541
41. Yang X, Chen Y, Fujita H, Liu D, Li T (2022) Mixed data-driven sequential three-way decision via subjective–objective dynamic fusion. Knowl-Based Syst 237:107728
42. Kostopoulos G, Karlos S, Kotsiantis S, Ragos O (2018) Semi-supervised regression: a recent review. J Intell Fuzzy Syst 35(2):1483–1500
43. Xu W, Guo Y (2016) Generalized multigranulation double-quantitative decision-theoretic rough set. Knowl-based Syst 105:190–205
44. Sang B, Yang L, Chen H, Xu W, Guo Y, Yuan Z (2019) Generalized multi-granulation double-quantitative decision-theoretic rough set of multi-source information system. Int J Approx Reason 115:157–179
45. Li W, Xu W, Zhang X, Zhang J (2021) Updating approximations with dynamic objects based on local multigranulation rough sets in ordered information systems. Artif Intell Rev 55:1821–1855
46. Zhou Z, Li M (2010) Semi-supervised learning by disagreement. Knowl Inf Syst 24(3):415–439
47. Triguero I, García S, Herrera F (2015) Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. Knowl Inf Syst 42(2):245–284
48. Tanha J, van Someren M, Afsarmanesh H (2017) Semi-supervised self-training for decision tree classifiers. Int J Mach Learn Cybernet 8(1):355–370
49. Zhang M, Zhou Z (2011) Cotrade: confident co-training with data editing. IEEE Trans Syst Man Cybernet Part B (Cybernet) 41(6):1612–1626
50. Eibe F, Hall MA, Witten IH (2016) The weka workbench. In: Online appendix for data mining: practical machine learning tools and techniques Morgan Kaufmann. Elsevier, Amsterdam
51. Sun L, Wang T, Ding W, Xu J, Lin Y (2021) Feature selection using fisher score and multilabel neighborhood rough sets for multilabel classification. Inf Sci 578:887–912
52. He X, Cai D, Niyogi P (2005) Laplacian score for feature selection, advances in nerual information processing systems, MIT Press, Cambridge