Information Sciences 580 (2021) 111–128

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Granular-conditional-entropy-based attribute reduction for partially labeled data with proxy labels

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

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

ARTICLE INFO

Article history:
Received 11 March 2021
Received in revised form 8 August 2021
Accepted 19 August 2021
Available online 22 August 2021

Keywords:
Rough sets
Semi-supervised attribute reduction
Conditional entropy
Information granularity
Proxy label

ABSTRACT

Attribute reduction is attracting considerable attention in the theory of rough sets, and thus many rough-set-based attribute reduction methods have been presented. However, most of them are specifically designed for either labeled or unlabeled data, whereas many real-world applications involve partial supervision. In this paper, we propose a rough-set-based semi-supervised attribute reduction method for partially labeled data. Specifically, using prior class-distribution information, we first develop a simple yet effective strategy to produce proxy labels for unlabeled data. Then, the concept of information granularity is integrated into an information-theoretic measure, based on which, a novel granular conditional entropy measure is proposed, and its monotonicity is theoretically proved. Furthermore, a fast heuristic algorithm is provided to generate the optimal reduct of partially labeled data, which could accelerate the process of attribute reduction by removing irrelevant examples and simultaneously excluding redundant attributes. Extensive experiments conducted on UCI data sets demonstrate that the proposed semi-supervised attribute reduction method is promising and, in terms of classification performance, it even compares favorably with supervised methods on labeled and unlabeled data with true labels (Our code and experimental data are released at Mendeley Data https://doi.org/10.17632/v3byhx2v8s.1).

1. Introduction

In many real-world applications, such as image classification, text mining, and gene analysis, the data to be processed are described by hundreds or thousands of attributes, posing a substantial challenge for conventional data analysis. Attribute reduction [12,31] (aka feature selection) has been proved to be effective in selecting the most informative attributes and removing irrelevant or redundant attributes from data. Attribute reduction has become an important preprocessing step in machine learning, pattern recognition, and data mining, as it enhances learning performance, increases computational efficiency, improves interpretability, and alleviates overfitting.

The theory of rough sets [23] is a representative granular computing [39,41] methodology for vague, uncertain, or imprecise data. Since the pioneering work of Pawlak [22], it has been rapidly developed both in theory and application [7,38,48].

[*] Corresponding author at: College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, PR China.
E-mail address: jie_jpu@163.com (J. Zhou).

https://doi.org/10.1016/j.ins.2021.08.067
0020-0255/© 2021 Elsevier Inc. All rights reserved.

Attribute reduction [31,8] is attracting considerable attention in the theory of rough sets. The objective of rough-set-based attribute reduction is to find an attribute subset with the same discriminability as the original attribute set. To evaluate the informativeness of attributes, one can use the positive region [21], the discernibility matrix [23], and information-theoretic methods [24]. Among them, information entropy is an efficient measure for certain and uncertain data, and thus many information-theoretic measures have been proposed for attribute reduction. Pawlak et al. [24] introduced conditional entropy to measure the significance of attributes. Liang and Qian [14,26] introduced the concepts of complement, rough, and combination entropy to measure uncertainty and fuzziness in rough set theory. Sun et al. [30] extended rough entropy to incomplete information systems. Jiang et al. [10] used relative decision entropy to select informative attributes from decision information system. Gao et al. [3,4] proposed maximum decision entropy and granular maximum decision entropy for attribute reduction. Zhang et al. [49] studied information entropy in fuzzy information system. Xu et al. [37] extended information entropy to fuzzy incomplete system. Wang et al. [32,33] defined several self-information measures and fuzzy extensions for attribute reduction in neighborhood information systems. Pawlak's rough set model relies on the indiscernibility relation, and thus tolerance to noise is limited. Owing to its superior generalizability and interpretability, three-way decision [13,25,36,40,42–47] has been extensively investigated, and many three-way-decision-based attribute reduction methods have been proposed, such as decision region distribution [18], minimum cost/risk [9], multi-scale decision[1], etc.

The aforementioned methods are often used for either labeled or unlabeled data. However, many real-world applications, such as web-page categorization, medical diagnosis, and defect detection [50], involve both labeled and unlabeled data. Therefore, semi-supervised attribute reduction based on rough sets is worthwhile to consider. To handle data with partial supervision (referred to as partially labeled data hereafter), the concepts of semi-supervised discernibility matrix [5,19] and discernibility pair [2] have been developed to yield a reduct of partially labeled data with categorical attributes, as the discernibility matrix can be used to extract discernible information of both labeled and unlabeled data. Instead of the equivalence relation for categorical data, neighborhood approximate quality [16], neighborhood decision error [17], and neighborhood granulation [11] have been proposed to handle partially labeled data with numerical attributes. To effectively handle large-scale partially labeled data, Qian et al. [27,28] first proposed the model of local rough sets and multigranulation extensions. Wang et al. [35] combined neighborhood rough sets and local rough sets, and developed the model of local neighborhood rough sets. Guo et al. [6] studied relative and absolute quantitative information in concept approximation, and presented the model of double-quantitative rough sets. Wang et al. [34] introduced the concept of local equivalence classes and proposed the model of double-local rough sets to improve computational efficiency. Additionally, many other rough-set-based methods have been proposed for semi-supervised classification [20] and semi-supervised clustering [15].

The aforementioned semi-supervised attribute reduction methods have their limitations. Specifically, to obtain the semi-supervised reduct, some of these methods use a carefully designed and complex mechanism to generate labels for unlabeled data, severely limiting their applicability to real-world tasks. Furthermore, the efficiency of the search algorithm is also an important factor in attribute reduction methods. However, existing methods should examine all candidate attributes in each iteration to find the optimal attribute, and thus the process of generating a semi-supervised reduct is highly time-consuming in the case of high-dimensional data. To address these limitations, we propose an effective semi-supervised attribute reduction method for partially labeled data. Unlike in the case of local rough sets and their extensions, the proposed method employs a prior-knowledge embedded-labeling strategy to assign proxy labels to unlabeled data rather than out-of-data class labels. Thus, in the process of attribute reduction, more informative and descriptive attributes can be selected to obtain the semi-supervised reduct. The main contributions of this study are as follows.

(1) To avoid a complex annotation mechanism for unlabeled data, a labeling strategy is designed whereby the class distribution of partially labeled data is regarded as prior knowledge and is used along with the distribution of the labeled data to determine proxy labels for the unlabeled data. This strategy is simple yet effective and has better applicability.
(2) To better evaluate the significance of attributes, a novel information-theoretic measure is proposed for attribute reduction. It incorporates information granularity with conditional entropy, and its monotonicity is theoretically proved.
(3) To expedite attribute reduction, we develop a strategy to accelerate the search algorithm by simultaneously excluding unnecessary examples and filtering redundant attributes. Moreover, extensive experiments are conducted to verify the effectiveness of the proposed model, and highly promising results are obtained.

The rest of the paper is organized as follows. In Section 2, we present preliminaries on rough sets and semi-supervised learning. In Section 3, we elaborate on the proposed semi-supervised attribute method for partially labeled data. The experiments and the related analysis are discussed in Section 4. Finally, Section 5 concludes the paper.

2. Preliminaries

Herein, we briefly review basic concepts related to rough sets and semi-supervised learning. More details can be found in [22,23,50].

2.1. Rough sets

In rough set theory, the data of interest are called an information system [23], which is denoted as $IS = (U, A)$; here, U is a set of examples, called the universe, and A is a set of attributes that describes the examples. The information system is also called a decision information system or decision table if $A = C \cup D$, where C is a set of condition attributes, and D is the decision attribute [23].

Given an attribute subset B of A, the universe U is partitioned into a family of equivalence classes U/B. The equivalence class containing x is denoted as $[x]_B$ and is referred to as a B-elementary granule [23]. Let X be a subset of the universe U. Then, the lower and upper approximations of X with respect to B are defined as follows [23]:

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

The B-lower approximation of X is the set of examples, the B-elementary granules of which belong to X, whereas the B-upper approximation of X is the set of examples, the B-elementary granules of which have a non-empty intersection with X. X is called a rough set with respect to B if $\underline{B}(X) \neq \overline{B}(X)$; otherwise X is a crisp set.

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

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

Let MES be a measure that quantifies the correlation between the condition attributes and the decision attribute. Then, for an attribute subset P of C, P is a reduct of C with respect to D if and only if [23].

(I) $MES_P(D) = MES_C(D)$, and.

(II) $\forall a \in P \wedge P^* = P - \{a\}, MES_{P^*}(D) \neq MES_C(D)$.

Condition (I) ensures that the data after attribute reduction have the same descriptive ability as the original data, and classification ability is thus preserved. Condition (II) is used to keep the attribute subset with the minimum redundancy. That is, each attribute in the reduct is individually necessary. In rough sets, the measure MES could be, for example, the positive region [23], information entropy [24], or discernibility preservation [23].

2.2. Semi-supervised learning

Semi-supervised learning is an efficient methodology for partially labeled data. Generally, the partially labeled data $PS = (U = L \cup N, A = C \cup D)$ are a combination of two sets of examples: the labeled set $L = \{x_i, y_i\}_{i=1}^{l}$ and the unlabeled set $N = \{x_i, ?\}_{i=l+1}^{u=l+n}$, where l is the number of labeled examples, n is the number of unlabeled examples, and $l \ll n$. In the context of semi-supervised learning, the label information of the labeled data can be used to enhance the results of unsupervised clustering (semi-supervised clustering). Moreover, the geometric structure of the unlabeled data can be captured to improve the performance of supervised methods trained only on the labeled data (Fig. 1), thereby obtaining semi-supervised attribute reduction, semi-supervised classification, or semi-supervised regression. The detailed description of these methods can be found in [50]. In this study, we only focus on semi-supervised attribute reduction.

In semi-supervised attribute reduction, a large amount of unlabeled data are employed to aid the selection of informative attributes when the labeled data at hand are scarce. As in traditional supervised attribute reduction, semi-supervised attribute reduction can be categorized into filter, wrapper, and embedded methods [29]. However, most existing methods apply to partially labeled data with numerical attributes, whereas partially labeled data with categorical attributes have received little attention.

3. Semi-supervised attribute reduction for partially labeled data

Herein, we first describe a strategy to generate proxy labels for unlabeled data. An improved information-theoretic measure is then developed, and a heuristic semi-supervised attribute reduction algorithm is proposed for partially labeled data with proxy labels.

3.1. Proxy label generation guided by prior knowledge

Traditional rough-set-based attribute reduction methods have been developed for either labeled or unlabeled data. In the case of partially labeled data, applying attribute reduction to the labeled data only may be insufficient owing to the presence

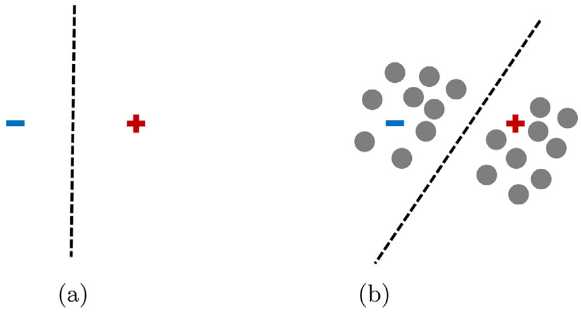


Fig. 1. Semi-supervised learning. (a) Decision boundary using only labeled data; (b) Decision boundary using both labeled and unlabeled data.

of a large amount of unlabeled data. Furthermore, unsupervised attribute reduction performed on the partially labeled data without labels results in the waste of valuable label information. Accordingly, it appears promising to use both labeled and unlabeled data to perform attribute reduction. In this study, we consider a strategy for annotating unlabeled data with proxy labels.

In practical semi-supervised applications, there is domain-specific knowledge that can be utilized to facilitate the learning process. For example, in the detection of lung cancer, a large number of medical images can be easily collected in routine diagnosis, but only a few representative images may be labeled by medical experts, as labeling all images is expensive and time-consuming. However, the occurrence probability of disease is generally known by domain experts in advance. Therefore, we could use this prior information to aid the learning of partially labeled data. More specifically, the distribution of different classes in practical tasks is regarded as the domain prior knowledge, whereas the distribution of initially labeled data in partially labeled data is used as the explicit intrinsic information. Then, the two types of information are integrated to determine proxy labels for unlabeled data in partially labeled data.

Formally, we assume that the partially labeled data $PS = (U = L \cup N, A = C \cup D)$ contain labeled data L with l examples, and unlabeled data N with n examples, where $u = l + n$ and $u = |U|$. Without loss of generality, we assume that the partially labeled data have two classes only, namely, we consider a binary classification problem. In the initially labeled data L, the sets of positive and negative examples are denoted as L_{pos} and L_{neg}, respectively, and the ratio of positive examples to negative examples is denoted as $\gamma = |L_{pos}|/|L_{neg}|$. The prior probability of positive examples over all examples is denoted as $P_{pos}(U) = |U_{pos}|/|U|$.

Considering the prior information $P_{pos}(U)$ regarding the partially labeled data and the class distribution of the initially labeled data, the unlabeled examples in the partially labeled data are assigned the proxy label y_{proxy} by the following formula:

$$y_{proxy} = \begin{cases} y_{pos}, & \lambda \leqslant 0.5 \\ y_{neg}, & \lambda > 0.5 \end{cases} \tag{3}$$

where $\lambda = P_{init}(\delta, \varepsilon) * \ P_{prior}(\varepsilon)$ and

$$P_{init}(\delta, \varepsilon) = \begin{cases} \gamma^{\left(1+e^{-\varepsilon\delta|L|}\right)}, & |L| \leqslant \delta \\ 1, & |L| > \delta \end{cases} \tag{4}$$

$$P_{prior}(\varepsilon) = \begin{cases} \min\left(P_{pos}(U) * (1+\varepsilon)^{|U|}, 0.5\right), & P_{pos}(U) \leqslant 0.5 \\ 1 - \min\left((1 - P_{pos}(U)) * (1+\varepsilon)^{|U|}, 0.5\right), & P_{pos}(U) > 0.5 \end{cases} \tag{5}$$

Formally, the determination of proxy labels for unlabeled data involves two correlated parts. The first is closely related to the distribution of the initially labeled data. When the amount of initially labeled data is small, the ratio γ of positive examples to negative examples influences greatly the determination of proxy labels. When $\gamma < 1$, the number of positive labeled examples is smaller than that of negative labeled examples. That is, the class distribution is imbalanced, adversely affecting the construction of a learning model. Therefore, the initial part P_{init} is used to balance the problem and assign positive proxy labels to unlabeled data. A smaller amount of initially labeled data implies a greater imbalance and a higher probability of positive labeling for the unlabeled data. Analogous considerations apply if $\gamma > 1$. If $\gamma = 1$, the distribution of different classes is relatively balanced. The learning model does not suffer from class imbalance, and this part does not need to be considered in the labeling strategy. However, the effect of γ gradually weakens as the amount of initially labeled data increases. Since the adverse effect of class imbalance could be partly negated by enlarging the size of the example set, we use a truncation function to suppress the effect of initially labeled data on the determination of proxy labels. Fig. 2 shows the effect of the first part on the labeling strategy for different parameters.

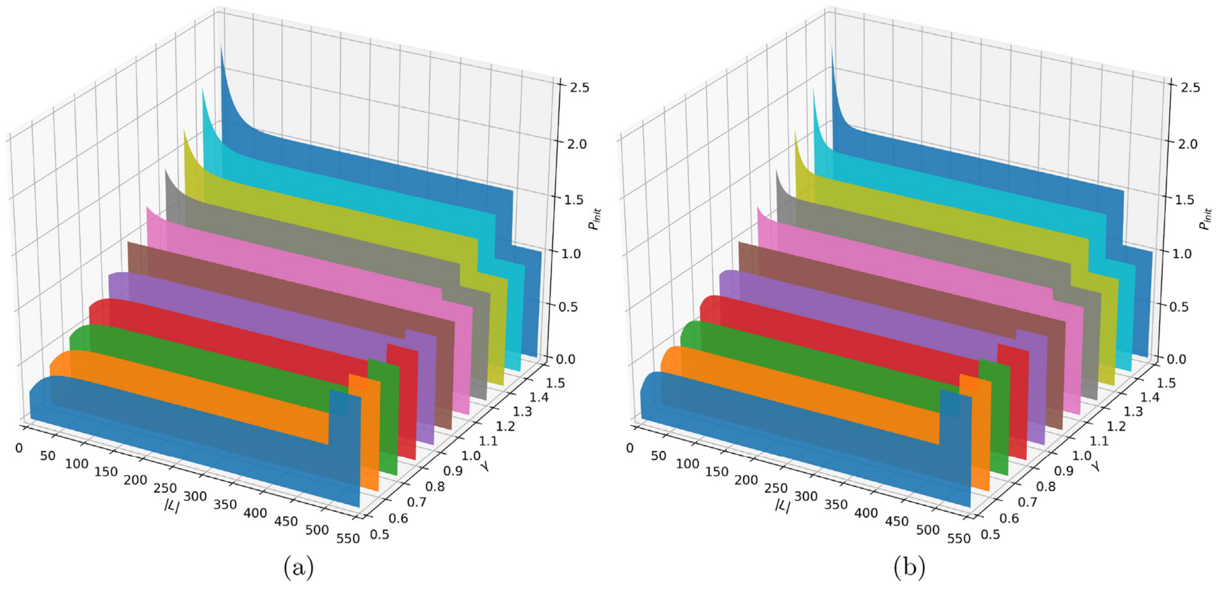


Fig. 2. The effect of class distribution of the initially labeled data on the labeling strategy. (a) P_{init} when $\delta = 500$ and $\varepsilon = 0.0001$; (b) P_{init} when $\delta = 500$ and $\varepsilon = 0.0002$.

The second part incorporates the combined effect of prior knowledge and data size. The prior probability is highly useful information in the assignment of proxy labels. If the prior probability of the positive class is smaller than 0.5, the initially labeled data contain fewer positive examples. Thus, the labeling strategy should assign positive proxy labels to unlabeled data to enrich the set of positive examples. Analogous considerations apply if the prior probability of the negative class is smaller than 0.5. Furthermore, data size is also of great importance. The labeling strategy relies heavily on the prior probability when data size is small. However, as the data size increases, the prior part P_{prior} tends to 0.5. That is, when data size is very large, the initially labeled data already contain a certain number of examples, and proxy labels are arbitrarily assigned. Fig. 3 shows the effect of the second part on the labeling strategy for different parameters.

The labeling strategy considers the data size $|U|$, the number of initially labeled examples $|L|$, the prior probability $P_{pos}(U)$, the class ratio γ, the boosting factor ε, and the truncation threshold δ. In fact, there are only two parameters ε and δ that should be set because the other parameters are automatically determined when the partially labeled data and prior knowledge are given. In Figs. 2 and 3, it can be observed that the effect of the initially labeled data weakens as the number of labeled examples increases, and the influence of data size strengthens with a slight increase in the parameter ε. In the following, these two parameters are empirically set to $\varepsilon = 0.0002$ and $\delta = 500$, respectively, to balance the effect of data size, prior probability, and initially labeled data on the labeling strategy.

3.2. Semi-supervised attribute reduction based on granular conditional entropy

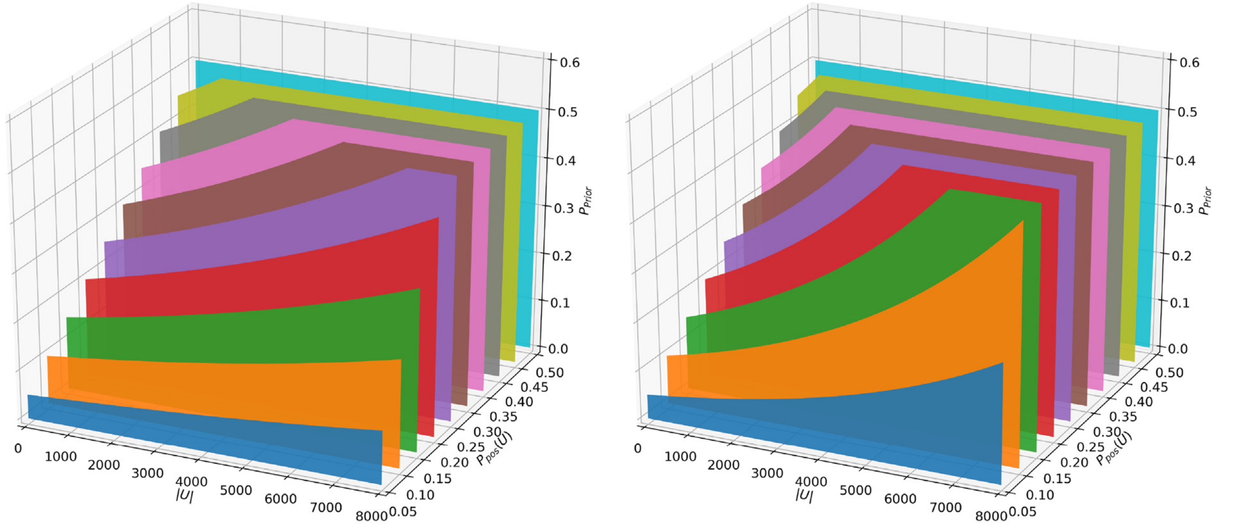
Information entropy is an efficient uncertainty measure, and thus it is often used to estimate the correlation or redundancy between attributes. In this study, we propose granular conditional entropy to evaluate the importance of attributes in partially labeled data with proxy labels. Formally, the partially labeled data after adopting the proposed labeling strategy are denoted as $PS = (U = L \cup N', A = C \cup D)$.

Definition 1. Let $PS = (U = L \cup N', A = C \cup D)$ be partially labeled data with proxy labels, and $U/B = \{X_1, X_2, \ldots, X_{|U/B|}\}$ be the partition induced by the condition attribute subset $B \subseteq C$. Then, the entropy of B over U is defined as follows [24]:

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

where $P(X_i) = |X_i|/|U|$and $|\cdot|$ denotes the cardinality of a finite set.

Definition 2. Let $PS = (U = L \cup N', A = C \cup D)$ be partially labeled data with proxy labels, and $U/B = \{X_1, X_2, \ldots, X_{|U/B|}\}$ and $U/D = \{Y_1, Y_2, \ldots, Y_{|U/D|}\}$ be the partitions induced by the condition attribute subset $B \subseteq C$ and the decision attribute D, respectively. Then, the conditional entropy of D given B is defined as follows [24]:



(a) (b)

Fig. 3. The effect of prior knowledge on the labeling strategy. (a) P_{prior} when $\varepsilon = 0.0001$; (b) P_{prior} when $\varepsilon = 0.0002$.

$$H(D|B) = -\sum_{i=1}^{|U/B|}\sum_{j=1}^{|U/D|} P(X_i, Y_j) \log P(Y_j|X_i), \tag{7}$$

where $P(X_i, Y_j) = P(Y_j|X_i)/P(X_i)$ and $P(Y_j|X_i) = |X_i \cap Y_j|/|X_i|$.

Definition 3. Let $PS = (U = L \cup N', A = C \cup D)$ be partially labeled data with proxy labels, and $U/B = \{X_1, X_2, \ldots, X_{|U/B|}\}$ be the partition induced by the condition attribute subset $B \subseteq C$. Then, the granularity of B over U is defined as follows [14]:

$$G(B) = -\sum_{i=1}^{|U/B|} P(X_i)^2. \tag{8}$$

Definition 4. Let $PS = (U = L \cup N', A = C \cup D)$ be partially labeled data with proxy labels, and $U/B = \{X_1, X_2, \ldots, X_{|U/B|}\}$ and $U/D = \{Y_1, Y_2, \ldots, Y_{|U/D|}\}$ be the partitions induced by the condition attribute subset $B \subseteq C$ and the decision attribute D, respectively. Then, the granular conditional entropy of D given B is defined as follows:

$$GH(D|B) = -\sum_{i=1}^{|U/B|} P(X_i)^2 \sum_{j=1}^{|U/D|} P(Y_j|X_i) \log P(Y_j|X_i). \tag{9}$$

For a given subset of condition attributes, the universe is partitioned into a family of condition classes. Conditional entropy formally accumulates the uncertainty of each condition class under different decisions. The partition induced by an attribute subset substantially reflects the discriminability of the attribute subset to a certain extent, and granularity can be used to evaluate the quality of the partition. Finer granularity implies stronger discriminating power. Accordingly, granular conditional entropy incorporates granularity with conditional entropy, thus providing a better measure for an attribute or attribute subset.

Proposition 1. *Let* $PS = (U = L \cup N', A = C \cup D)$ *be partially labeled data with proxy labels. Then, for any attribute subset* $B \subseteq C, 0 \leqslant GH(D|B) \leqslant \log |U|$.

Proof. Granular conditional entropy is minimized when the decision attribute D is completely dependent on the condition attribute subset B. That is, each equivalence class induced by the condition attribute subset B has only one decision, the conditional probability of D given B is 0, and the overall granular conditional entropy is thus minimized to 0. The granular conditional entropy is maximized when the decision attribute D is conditionally independent of the condition attribute subset B, namely $GH(D|B) = H(D)$, whereas $H(D)$ takes the maximum value $\log |U|$ when the probability distribution is uniform. Thus, $GH(D|B) \leqslant \log |U|$. The proposition is proved.

Proposition 2. *Let $PS = (U = L \cup N', A = C \cup D)$ be partially labeled data with proxy labels, and $P,\ Q$ be subsets of C. If $P \subset Q$, then $GH(D|P) \geqslant GH(D|Q)$.*

Proof. Without loss of generality, we assume that $Q = P \cup \{a\}$, and only the equivalence class X_{ij} under P is divided into the equivalence classes X_i and X_j under Q after the attribute a is added. Namely, $U/P = \{X_1, X_2, \ldots, X_{ij}, \ldots, X_n\}$ and $U/Q = \{X_1, X_2, \ldots, X_i, X_j, \ldots, X_n\}$. The proof is presented in the appendix.

Proposition 2 implies that granular conditional entropy decreases monotonically as attributes are added, and can thus be considered a measure of attribute reduction.

Definition 5. Let $PS = (U = L \cup N', A = C \cup D)$ be partially labeled data with proxy labels, and let $P \subset C$. Then, for a condition attribute $a \in (C - P)$, the relative significance of D given P is defined as follows:

$$Sig(a, P, D) = GH(D|P) - GH(D|(P \cup \{a\})). \tag{10}$$

Definition 6. Let $PS = (U = L \cup N', A = C \cup D)$ be partially labeled data with proxy labels. Then, an attribute subset P of C is a reduct of C with respect to D if and only if

(I) $GH(D|P) = GH(D|C)$, and

(II) $\forall a \in P \wedge P^* = P - \{a\}, GH(D|P^*) \neq GH(D|C)$.

Algorithm 1. Accelerated semi-supervised attribute reduction algorithm based on granular conditional entropy.

```
Input: Partially labeled data $PS = (U = L \cup N, A = C \cup D)$, prior probability of positive
class $P_{pos}(U)$, and threshold parameters $\delta$ and $\varepsilon$.
Output: Optimal semi-supervised reduct $RED$;
1: Compute the class ratio $\gamma$ within the initially labeled data $L$;
2: Determine the proxy labels of unlabeled data $N$ using the prior probability $P_{pos}(U)$, the class ratio $\gamma$, and the threshold
   parameters $\delta$ and $\varepsilon$;//refers to Formula (3)
3: Compute the overall granular conditional entropy $GH(D|C)$;
4: Evaluate each attribute using the granular conditional entropy $GH(D|\{a_i\})$, and add the attribute
   $a_{opt} = \mathrm{argmin}_{a_i \in C}\{GH(D|\{a_i\})\}$ to $RED$;
5: While $GH(D|RED) \neq GH(D|C)$ Do
6:     Compute the relative significance of each attribute $a_i$ for $D$ given $RED$ and the granular conditional entropy
       $GH(\{a_i\}|RED)$;
7:     Select an attribute $a_{opt}$ with maximum significance and remove the attributes with $GH(\{a_i\}|RED) = 0$;//
       Acceleration in attribute
8:     $RED \leftarrow RED \cup \{a_{opt}\}$ and remove examples that their granular conditional entropy under $RED$ is 0;//Acceleration in
       example
9: End While
10: Return Semi-supervised reduct $RED$.
```

Attribute reduction is strongly related to the measure for evaluating the significance of attributes, and involves the strategy of finding the reduct. It is well known that finding the minimum reduct or all reducts is NP-hard, and thus a heuristic method is preferred. Existing heuristic methods can be divided into forward adding, backward deleting, and bi-directional adding–deleting. Considering efficiency, we use the strategy of forward adding to perform attribute reduction for partially labeled data.

Using the proposed granular conditional entropy, we design two tactics to accelerate attribute reduction. Specifically, granular conditional entropy decreases monotonically as attributes are added to the reduct. During the process of attribute reduction, if a condition equivalence class has zero granular conditional entropy in one stage, its granular conditional entropy is always zero in the following stages. Therefore, such examples in the condition equivalence class can be removed without further consideration. Furthermore, to find informative attributes for the reduct, the attribute reduction algorithm examines the relative significance of each candidate attribute, and selects the optimal attributes. Considering that attributes are correlated, and there may be some attributes such that their granular conditional entropy with respect to the selected optimal attributes is zero, implying that these attributes are redundant with the decision attribute given the selected attributes, and thus they can also be excluded from the list of candidate attributes for the reduct. The overall attribute reduction algorithm based on granular conditional entropy, incorporating the acceleration strategy, is presented Algorithm 1.

The algorithm first labels all unlabeled examples with the proxy labels determined by the prior probability and the class distribution of the initially labeled data (lines 1 and 2). The overall granular conditional entropy under all conditional attributes is then computed and is used as a stopping condition for the algorithm. In the first round of attribute selection, the algorithm evaluates each attribute using only its granular conditional entropy with respect to the decision attribute, and the attribute with minimal granular conditional entropy is selected as the optimal attribute. In the following rounds, the algorithm iteratively adds the optimal attributes with maximum significance to the reduct until the granular conditional entropy of the selected attributes reaches the stopping value (lines 5–9). Two acceleration strategies are incorporated into the algorithm, resulting in higher efficiency.

The main cost of Algorithm 1 is the iterative selection of the optimal attributes. We assume that a partially labeled data have $|U|$ examples described by $|C|$ attributes. In each iteration, the time cost for determining an optimal attribute is $O\left(|C||U|^2\right)$. In the worst case, the algorithm is terminated after $|C|$ rounds of selection. Therefore, the time cost for computing an optimal reduct is at most $O\left(|C|^2|U|^2\right)$, and the total space cost is at most $O(|C||U|)$. Considering the acceleration strategy, the overall cost of Algorithm 1 is considerably lower in terms of both time and space.

4. Empirical analysis

Herein, we first evaluate the effectiveness of the proposed method. Then, we compare the proposed method with other classic methods in terms of classification accuracy. All methods were implemented in Python 3.6 on the PyCharm development platform. All experiments were conducted on a Windows 10 PC with Intel Xeon (R) CPU E5-2650 v4@2.20 GHz processor and 128 GB RAM.

4.1. Data sets and experiment design

Twelve UCI data sets[1] were used in the experiments, and the details are shown in Table 1. We note that some of the data sets are multiclass, thus the “1-vs-rest” criterion was employed to obtain two-class data. More specifically, the class with the largest number of examples was considered the positive class, and the other classes were grouped into the negative class.

In Table 1, the number of examples is shown in the second column. The number of condition attributes is presented in the third column, where the number of numerical attributes is also listed in the brackets. The number of classes in the original data set is given in the fourth column. The last column indicates the class distribution after the “1-vs-rest” criterion was applied. This distribution is the prior knowledge that is used to guide the determination of proxy labels.

In the experiments, all numerical attributes were discretized into categorical ones using the principle of equal frequency binning with three bins. More specifically, all examples were first ranked in ascending order according to their values, and then divided into three bins, each of which had the same number of examples. In real-world applications, the number of labeled examples is usually very small. To fully examine the effectiveness of the proposed methods on partially labeled data, the label rate α was set in the range [0.01, 0.3] in the experiments. For a given label rate α, each data set was first partitioned into a set of labeled examples L and a set of unlabeled data N. Then, the prior class probability and the class ratio of the initially labeled data L were used to determine the proxy labels of the unlabeled examples N. To gain a deeper insight into the labeling strategy, we used data partitions with different positive ratios. More specifically, for a given label rate α, the positive ratio β of the initially labeled data varied from 0.5 to 1.5. For example, given a partially labeled data set with 1000 examples, prior probability $P_{pos}(U) = 0.5$, and label rate $\alpha = 10\%$, a labeled set L with 25 positive examples $\left(|L_{pos}| = \beta * P_{pos}(U) * \alpha * |U|\right)$ and 75 negative examples $\left(|L_{neg}| = \alpha * |U| - |L_{pos}|\right)$ was randomly generated if $\beta = 0.5$, and then the remaining 900 examples were grouped into a set of unlabeled examples. If the label rate α, the positive ratio β, and the prior probability $P_{pos}(U)$ are small, $|L_{pos}|$ may be close to 0. To avoid a lack of examples in the positive class, we set $|L_{pos}|$ to 1. To ensure the validity of the results, we repeated the data partitioning 10 times for each pair of α and β, and the performance was averaged.

4.2. Attribute reduction for partially labeled data

To evaluate the effectiveness of the proposed algorithm, we conducted an experiment on all examples of each data set, with $\alpha = 10\%$ and β varying from 0.5 to 1.5. The reduct information is shown in Table 2.

In the table, statistical information, including the minimum, the maximum, and the average number of attributes in the obtained reducts for different positive ratios, is provided in the third to eighth columns. Moreover, the ground-truth reduct information is provided for comparison, namely the optimal reduct for a label rate $\alpha = 100\%$. The tenth column indicates the number of common attributes between the first semi-supervised reduct with the smallest number of attributes for different values of β and the ground-truth reduct, and the eleventh column indicates the similarity degree between the semi-supervised reduct and the ground-truth reduct, which is computed by dividing the value of “Num. of common attributes” by that of “Ground-truth”. The last column “Approx. rate” indicates the approximate degree between the semi-supervised

[1] http://archive.ics.uci.edu/ml/index.php.

Table 1
The experimental data sets.

Data set	$\|U\|$	$\|C\|$	$\|U/D\|$	$(P_{pos}(U), P_{neg}(U))$
cardiotocography-FHR pattern(cardio)	2126	21(21)	10	(0.2723, 0.7277)
frogs calls-genus(frog)	7195	22(22)	8	(0.5768, 0.4232)
gesture-phase-a3va3(gesture1)	1830	32(32)	5	(0.3595, 0.6405)
gesture-phase-b1va3(gesture2)	1069	32(32)	5	(0.3854, 0.6146)
kdd-synthetic–control(kdd)	600	60(0)	6	(0.1667, 0.8333)
kr-vs-kp(krvskp)	3196	36(0)	2	(0.5222, 0.4778)
landsat(landsat)	6435	36(0)	6	(0.2382, 0.7618)
libras movement(libras)	360	90(90)	15	(0.0667, 0.9333)
musk2(musk)	6598	166(0)	2	(0.8458, 0.1542)
spambase(spam)	4601	57(57)	2	(0.6060, 0.3940)
vehicle(vehicle)	846	18(18)	4	(0.2577, 0.7423)
wine(wine)	178	13(13)	3	(0.3989, 0.6011)
Avg.	2919.50	48.58(23.75)	5.67	(0.3914, 0.6086)

Table 2
The results of attribute reduction on the selected data sets (label rate $\alpha = 10\%$).

Data set	Raw	Only labeled data			Ours			Ground-truth	Num. of common attributes	Sim. rate	Approx. rate
		Min	Max	Avg.	Min	Max	Avg.				
cardio	21	6	11	8.04	17	20	19.19	17	15	0.88	0.89
frog	22	7	11	8.85	21	22	21.32	14	14	1.00	0.66
gesture1	32	7	13	9.21	19	27	22.35	21	17	0.81	0.94
gesture2	32	5	8	6.33	12	21	17.05	16	7	0.44	0.94
kdd	60	2	4	2.55	8	10	8.81	5	3	0.60	0.57
krvskp	36	10	18	13.59	29	33	31.49	29	28	0.97	0.92
landsat	36	7	18	12.18	36	36	36.00	29	29	1.00	0.81
libras	90	2	3	2.25	10	17	12.73	5	2	0.40	0.39
musk	166	2	14	8.07	37	72	57.86	24	3	0.13	0.41
spam	57	7	18	12.9	33	48	42.66	38	25	0.66	0.89
vehicle	18	3	7	4.90	12	16	14.09	10	8	0.80	0.71
wine	13	1	3	1.89	6	10	7.51	5	3	0.60	0.67
Avg.	48.58	4.92	10.67	7.56	20.00	27.67	24.26	17.75	12.83	0.69	0.73

reduct and the ground-truth reduct, which is the ratio of the value of “Ground-truth” to that of “Avg.” in the proposed method.

By observing the experimental results, we find that the class distribution of the initially labeled data has a great influence on the obtained reduct of the partially labeled data. A more balanced class distribution results in fewer attributes in the reduct. Moreover, completely irrelevant attributes are always excluded from the obtained reducts, whereas different weakly relevant attributes are removed from the reducts if different labeled examples are available. The reducts obtained only from labeled data appear to attain the best attribute reduction rate, but they could only discern the labeled data rather than the entire partially labeled data set, and thus they are not suitable for classification, as confirmed by the following experiments. Unlike supervised methods learned from labeled data only, the proposed method considers both labeled and unlabeled data. Thus, more attributes are selected by the proposed semi-supervised methods to ensure the discriminability of unlabeled data with proxy labels. On all data sets, the proposed method achieved a reduction rate of 50.08% over raw data, a similarity rate of 68.98%, and an approximate rate of 73.18% with respect to the ground-truth. It is worth mentioning that, on the data sets “frog,” “krvskp,” and “landsat”, the similarity rate is close to 1, implying that the obtained semi-supervised reduct has almost the same attributes as the ground-truth reduct. Moreover, on the data sets “cardio,” “gesture1,” and “krvskp”, the minimum number of attributes in the semi-supervised reduct is equal to or smaller than that of the ground truth. These results demonstrate the potential of the proposed method.

4.3. Effectiveness of the proposed method

To evaluate the quality of the reducts obtained by the proposed method, we conducted further experiments for a given label rate. Specifically, a supervised or semi-supervised reduct was first generated with α = 10%, and $\beta \in [0.5, 1.5]$. Redundant attributes not in the obtained reduct were then removed from each data set, and 10-fold cross-validation was performed on the reduced data set to evaluate the performance of the method. In the experiments, we used the following classifiers: k-nearest neighbor (KNN) with $k = 3$, and support vector machine (SVM) with radial basis function. The results are shown in Tables 3 and 4, respectively. Note that we shuffled each data set 10 times and performed 10-fold cross-validation on the shuffled data so that the order of the samples may not affect performance.

Table 3
The performance of the proposed method for different positive ratios (KNN with a label rate α = 10%).

Data set	0.5		0.6		0.7		0.8		0.9		1.0		1.1		1.2		1.3		1.4		1.5	
	initial	final	initial	final	initial	final	initial	final	initial	final	initial	final	initial	final	initial	final	initial	final	initial	final	initial	final
cardio	0.8557	0.8861	0.8539	**0.8876**	0.8591	0.8874	0.8601	0.8863	0.8679	0.8863	0.8622	0.8871	0.8615	0.8860	0.8637	0.8869	0.8674	0.8868	0.8660	0.8867	0.8611	0.8869
frog	0.9679	0.9862	0.9696	**0.9863**	0.9706	0.9863	0.9707	0.9861	0.9696	0.9857	0.9681	0.9858	0.9667	0.9859	0.9645	0.9858	0.9654	0.9858	0.9663	0.9858	0.9645	0.9858
gesture1	0.7899	0.8528	0.7996	0.8570	0.8035	0.8563	0.8173	0.8560	0.8066	0.8507	0.8086	0.8512	0.8015	0.8528	0.8045	**0.8610**	0.8067	0.8524	0.8117	0.8592	0.8076	0.8570
gesture2	0.7009	**0.7845**	0.7033	0.7711	0.6900	0.7712	0.6985	0.7780	0.6966	0.7662	0.6942	0.7670	0.6955	0.7734	0.6952	0.7659	0.6976	0.7723	0.6931	0.7455	0.6919	0.7575
kdd	0.9048	0.9815	0.9203	0.9783	0.9412	0.9760	0.9427	0.9752	0.9443	0.9808	0.9230	0.9783	0.9212	0.9743	0.9385	**0.9833**	0.9357	0.9782	0.9305	0.9817	0.9325	0.9793
krvskp	0.9344	0.9473	0.9308	0.9479	0.9299	0.9473	0.9329	0.9464	0.9298	0.9460	0.9355	0.9449	0.9311	**0.9493**	0.9349	0.9489	0.9333	0.9493	0.9400	0.9479	0.9283	0.9491
landsat	0.9501	0.9769	0.9571	0.9769	0.9582	0.9769	0.9557	0.9766	0.9604	0.9770	0.9607	0.9771	0.9605	**0.9772**	0.9609	0.9770	0.9594	0.9770	0.9623	0.9771	0.9633	0.9767
libras	0.8603	0.9672	0.8997	0.9664	0.8475	0.9608	0.8658	0.9664	0.8594	0.9611	0.8617	0.9631	0.8800	**0.9714**	0.8875	0.9625	0.8903	0.9617	0.8667	0.9614	0.8389	0.9683
musk	0.9179	0.9453	0.9239	0.9492	0.9228	0.9509	0.9249	0.9490	0.9174	**0.9517**	0.9178	0.9511	0.9050	0.9508	0.7661	0.9502	0.7729	0.9513	0.7956	0.9509	0.8063	0.9513
spam	0.8911	0.9249	0.8854	0.9228	0.8894	0.9189	0.8940	0.9206	0.8853	0.9249	0.8911	0.9229	0.8876	0.9218	0.8860	0.9227	0.8782	0.9192	0.8835	0.9221	0.8633	**0.9251**
vehicle	0.9123	0.9490	0.9019	0.9507	0.8903	0.9476	0.9126	0.9491	0.9115	0.9450	0.9079	0.9455	0.9214	0.9472	0.9211	**0.9512**	0.9123	0.9473	0.9119	0.9475	0.9084	0.9455
wine	0.7551	0.9029	0.7999	0.9104	0.8266	0.9067	0.8146	0.9234	0.7697	0.9133	0.8425	0.9178	0.7845	**0.9239**	0.7616	0.9166	0.8016	0.9234	0.8249	0.9126	0.8094	0.8729
Avg.	0.8700	0.9254	0.8788	0.9254	0.8774	0.9239	0.8825	0.9261	0.8766	0.9241	0.8811	0.9243	0.8764	**0.9262**	0.8654	0.9260	0.8684	0.9254	0.8710	0.9232	0.8646	0.9213

Table 4
The performance of the proposed method for different positive ratios (SVM with a label rate α = 10%).

Data set	0.5		0.6		0.7		0.8		0.9		1.0		1.1		1.2		1.3		1.4		1.5	
	initial	final	initial	final	initial	final	initial	final	initial	final	initial	final	initial	final	initial	final	initial	final	initial	final	initial	final
cardio	0.8306	0.8697	0.8374	0.8744	0.8438	0.8702	0.8475	0.8743	0.8502	0.8738	0.8465	0.8721	0.8444	0.8672	0.8490	0.8744	0.8455	**0.8753**	0.8421	0.8717	0.8398	0.8676
frog	0.9313	0.9486	0.9314	**0.9486**	0.9368	0.9485	0.9349	0.9484	0.9368	0.9482	0.9310	0.9482	0.9344	0.9482	0.9315	0.9482	0.9279	0.9482	0.9302	0.9482	0.9329	0.9482
gesture1	0.7322	0.7570	0.7367	0.7603	0.7314	0.7589	0.7308	0.7570	0.7374	0.7563	0.7365	0.7480	0.7328	0.7580	0.7411	0.7548	0.7290	0.7541	0.7478	**0.7657**	0.7433	0.7674
gesture2	0.6458	0.6522	0.6355	0.6504	0.6373	0.6542	0.6351	0.6492	0.6473	0.6488	0.6261	0.6500	0.6481	0.6622	0.6475	0.6549	0.6460	**0.6625**	0.6314	0.6485	0.6387	0.6504
kdd	0.9262	0.9803	0.9167	0.9765	0.9345	0.9772	0.9530	0.9813	0.9392	**0.9833**	0.9433	0.9755	0.9342	0.9712	0.9520	0.9818	0.9432	0.9753	0.9397	0.9823	0.9420	0.9780
krvskp	0.9465	0.9587	0.9445	0.9588	0.9479	0.9585	0.9483	0.9586	0.9467	0.9585	0.9472	0.9589	0.9462	0.9589	0.9485	0.9589	0.9450	**0.9592**	0.9473	0.9588	0.9465	0.9591
landsat	0.9299	0.9663	0.9334	0.9663	0.9407	0.9663	0.9369	0.9663	0.9410	0.9663	0.9397	0.9663	0.9372	0.9663	0.9374	0.9663	0.9357	**0.9663**	0.9392	0.9663	0.9435	0.9663
libras	0.9333	0.9547	0.9333	0.9508	0.9333	0.9494	0.9333	0.9489	0.9331	0.9514	0.9333	**0.9597**	0.9389	0.9589	0.9333	0.9533	0.9333	0.9531	0.9344	0.9467	0.9333	0.9589
musk	0.8572	0.9007	0.8591	0.9074	0.8626	0.9134	0.8547	0.9092	0.8607	0.9157	0.8548	0.9149	0.8528	0.9141	0.8459	**0.9176**	0.8459	0.9161	0.8459	0.9159	0.8459	0.9122
spam	0.8918	**0.9403**	0.8912	0.9380	0.8952	0.9375	0.8988	0.9388	0.8942	0.9395	0.8937	0.9389	0.8936	0.9401	0.8901	0.9394	0.8850	0.9395	0.8823	0.9400	0.8673	0.9393
vehicle	0.8561	0.9119	0.8519	0.9122	0.8266	0.9101	0.8353	**0.9216**	0.8476	0.9091	0.8488	0.9059	0.8531	0.9166	0.8544	0.9197	0.8454	0.9082	0.8391	0.9098	0.8325	0.8983
wine	0.8461	0.9125	0.8388	0.9170	0.8068	0.9276	0.8406	0.9315	0.8971	0.9193	0.8875	0.9249	0.8478	0.9292	0.8613	**0.9327**	0.8380	0.9276	0.8470	0.9125	0.7989	0.8622
Avg.	0.8606	0.8961	0.8592	0.8967	0.8581	0.8976	0.8624	0.8988	0.8693	0.8975	0.8657	0.8969	0.8636	0.8992	0.8660	**0.9002**	0.8600	0.8988	0.8605	0.8974	0.8554	0.8923

Tables 3 and 4 present the average performance over 10 runs of 10-fold cross-validation. For each positive ratio β, the columns "initial" and "final" respectively indicate the performance of the reduct obtained from the initially labeled data and that of the semi-supervised reduct further refined by unlabeled data with proxy labels. The highest performance for different positive ratios is in bold, and the average performance over all data sets is shown in the last row ("Avg.").

It can be seen that the quality of attribute reduction is significantly improved by unlabeled data. Owing to the scarcity of label information in partially labeled data, attribute evaluation using labeled data only may not reflect the importance of attributes, so that the supervised attribute reduction method generates low-quality reducts with few attributes, resulting in mediocre performance. Essentially, the semi-supervised reduct considers both labeled and unlabeled data. Moreover, attribute selection is guided by the proxy labels of unlabeled data, which is jointly determined by prior knowledge and the class information of the initially labeled data. As a result, the selected attributes are more representative and informative, and higher performance is achieved by the resulting semi-supervised reduct. On each data set, the supervised reduct achieved different performance when the positive ratio varied from 0.5 to 1.5, whereas the obtained semi-supervised reduct achieved relatively stable and high performance. This may be because the balance of the class distribution has a significant effect on performance, and the semi-supervised reduct could weaken this adverse effect by utilizing the proxy labels of unlabeled data. By averaging all results over different data sets, the proposed method using KNN and SVM achieved a maximum improvement of 7.01% ($\beta = 1.2$) and 4.61% ($\beta = 0.7$), respectively, over the supervised method. Interestingly, the proposed method achieved the highest performance when $\beta = 1.1$ (KNN) and $\beta = 1.2$ (SVM), for which the class ratio of the initially labeled data is close to 0.5 $\left(\beta * P_{pos}(U)\right)$. That is, the proposed method is likely to achieve the highest performance when the class distribution of the initially labeled data is balanced. These results clearly indicate that the proposed method is effective and could benefit from the proxy labels of unlabeled data to improve the quality of attribute reduction.

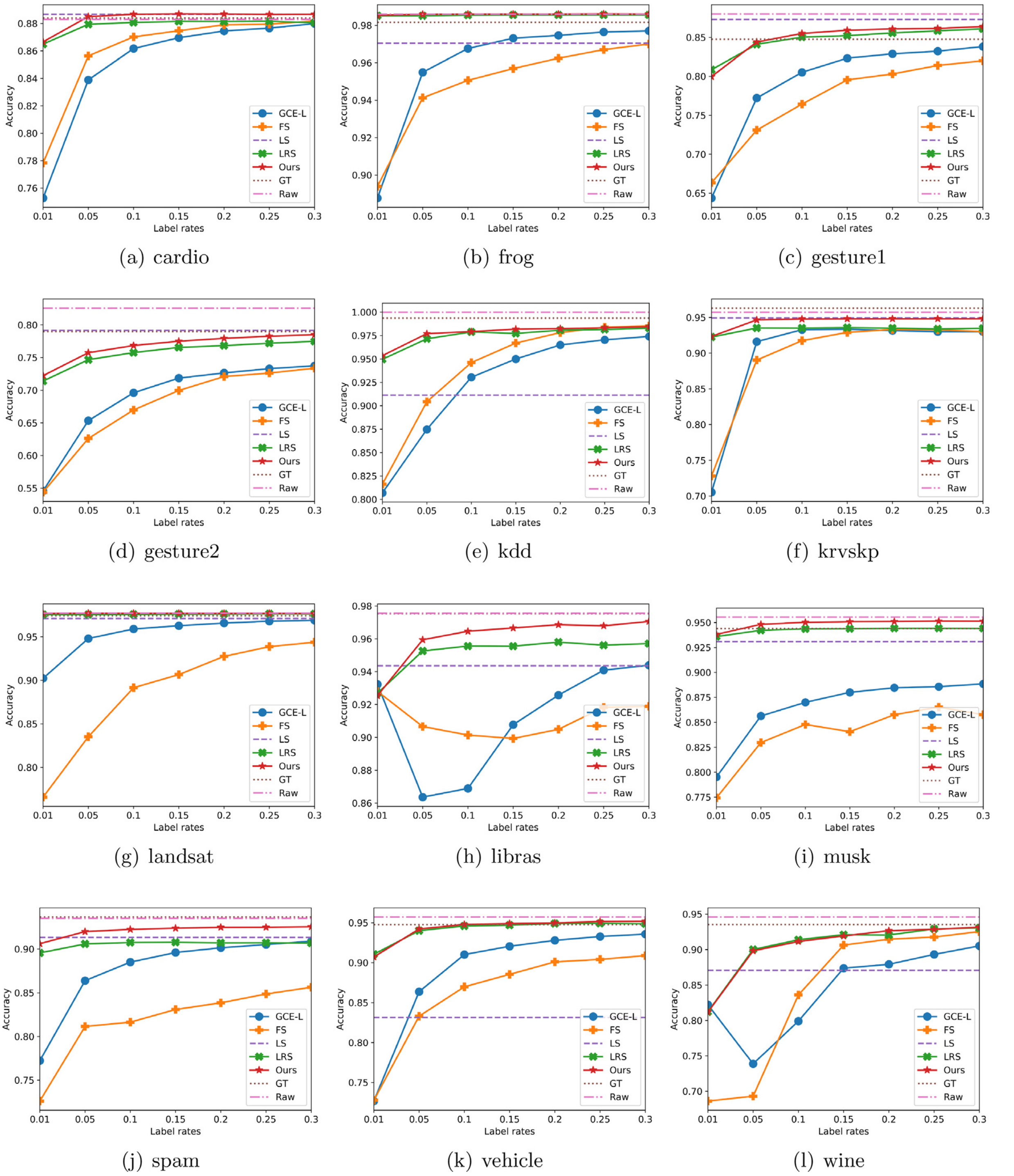
To further verify its effectiveness, the proposed method was compared with the following attribute reduction methods: supervised Fisher score (FS) [12], unsupervised Laplacian score (LS) [12], local rough sets on labeled and unlabeled data with pseudo-labels [28,34], the proposed granular conditional entropy only on labeled data (GCE-L), the proposed granular conditional entropy on all data with true labels (GT), and raw data without attribute reduction (Raw). In the semi-supervised method of local rough sets (LRS), the unlabeled data were assigned a pseudo-label as performed in [34], and the threshold for the lower approximate was set to 1. The other settings were the same as in the proposed method, namely the positive ratio β of initially labeled data ranged from 0.5 to 1.5 for each label rate, and the final performance was averaged over 10 runs of 10-fold cross-validation. It is observed that LRS usually selects more attributes to discern unlabeled data with out-of-data pseudo-labels. For a fair comparison, when the number of attributes was greater than the maximum number of attributes in the proposed method, the attribute subsets obtained by LRS were truncated by removing the last added attributes. The overall settings for all selected methods are shown in Table 5. The experiments were conducted for different label rates, and the results are shown in Figs. 4 and 5. Note that the performance of GCE-L, FS, LRS, and our method was averaged over different positive ratios.

It can be seen that the proposed method significantly outperforms the supervised methods with only initially labeled data on almost all data sets. Although FS and GCE-L are effective measures for attribute reduction, important attributes evaluated using labeled data only are not necessarily informative with respect to the entire partially labeled data set, resulting in poor performance. Surprisingly, the performance of the supervised methods with only labeled data appears rather unstable on some data sets, such as "libras" and "wine," on which methods with higher label rates have worse performance. This inconsistency may be attributed to the size of the data set and the class distribution of the initial labeled data. When the data set is small ("libras" has only 360 examples, and "wine" has 178 examples) and the class distribution of the initial labeled data is imbalanced, the quality of labeled data is undoubtedly poor, and thus even if the label rate increases, the performance of the supervised methods remains unstable.

Laplacian score is an effective unsupervised attribute reduction method utilizing all examples in partially labeled data. On the data sets "gesture1" and "gesture2," LS achieved slightly better performance than the proposed method. However, on the other data sets, the performance of LS is considerably worse than that not only of the proposed methods but also of the

Table 5
The settings for all selected methods.

Method	Attribute evaluation	Attribute subset
Granular conditional entropy on labeled data only (GCE-L)	$Sig(a, P, D)$	Attribute reduction
Fisher score on labeled data only (FS)	$F_{score}(a) = S_B(a)/S_W(a)$	Top k attributes
Laplacian score on all data without labels (LS)	$L_{score}(a) = \sum_{ij}(a(i) - a(j))^2 S_{ij}/Var(a)$	Top k attributes
Local rough sets on labeled data and unlabeled data with pseudo labels (LRS)	$Sig(a, P, D)$	Attribute reduction
Granular conditional entropy on labeled data and unlabeled data with proxy labels (Ours)	$Sig(a, P, D)$	Attribute reduction
Granular conditional entropy on all data with true labels (GT)	$Sig(a, P, D)$	Attribute reduction
All data without attribute reduction (Raw)	–	–



(a) cardio (b) frog (c) gesture1

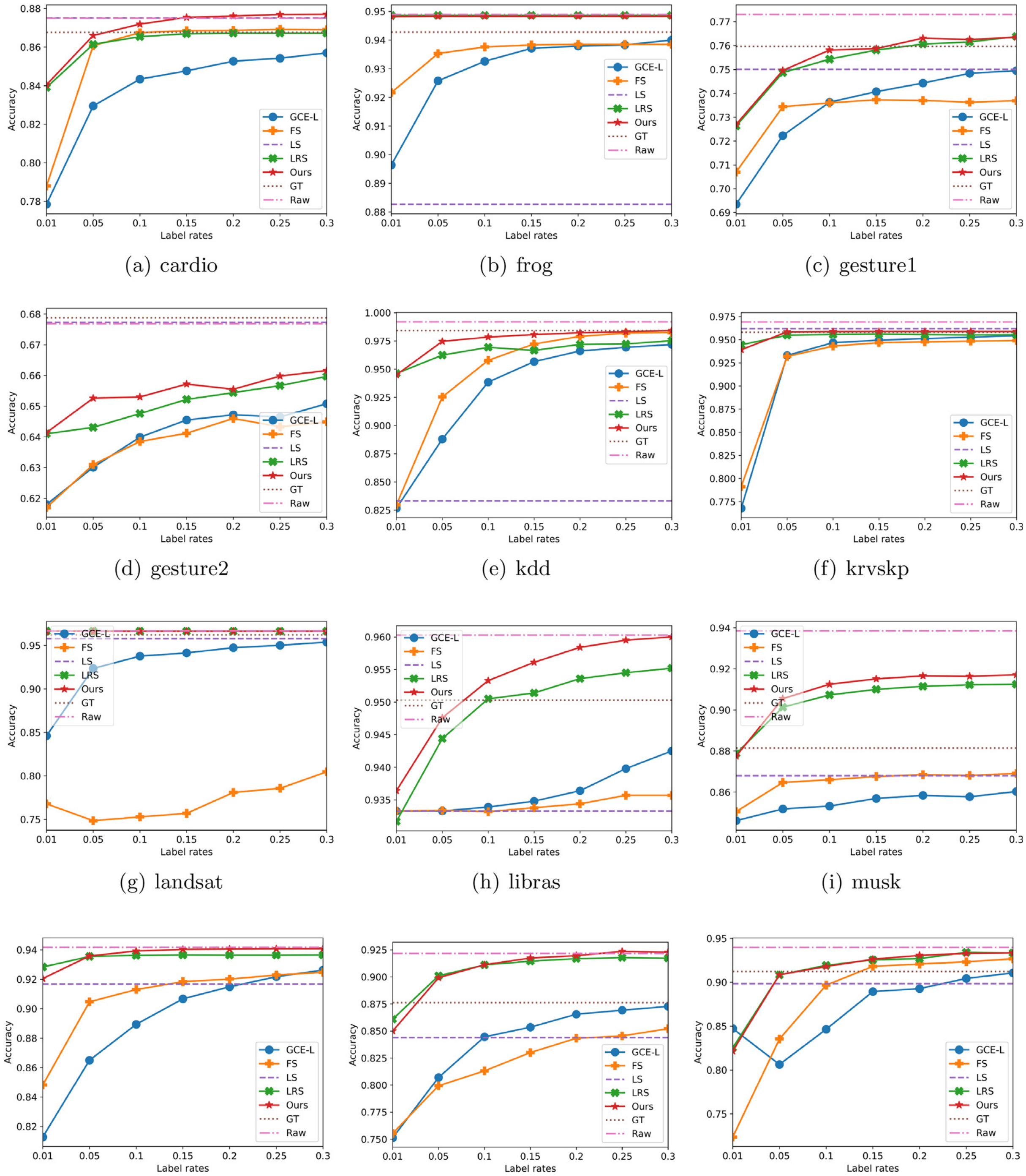
(d) gesture2 (e) kdd (f) krvskp

(g) landsat (h) libras (i) musk

(j) spam (k) vehicle (l) wine

Fig. 4. The performance of the selected methods for different label rates (KNN).

selected supervised method with labeled data only. LRS is a semi-supervised method and can use labeled and unlabeled data for attribute reduction. It achieved better performance than the supervised GCE-L and FS on most data sets. However, LRS labels the unlabeled data with a pseudo-class that is different from the real class of the labeled data, so that the significance of the attributes may not be well measured. Moreover, it discerns each unlabeled example from all labeled examples even if the unlabeled examples may have the same class value as the labeled ones. Thus, more attributes are selected in the process of attribute reduction. As a result, on the data sets “krvskp” and “spam” with KNN, and “cardio” and “kdd” with SVM, the



(a) cardio (b) frog (c) gesture1

(d) gesture2 (e) kdd (f) krvskp

(g) landsat (h) libras (i) musk

(j) spam (k) vehicle (l) wine

Fig. 5. The performance of the selected methods for different label rates (SVM).

performance of LRS is worse than that of the supervised methods for high label rates. The proposed method (Ours) uses both labeled and unlabeled data, and the latter are labeled with proxy labels that are determined by carefully considering the prior information of the entire data set and the class distribution of the initially labeled data. Moreover, the proposed measure for attribute reduction combines conditional entropy and information granularity, and a reduct with higher discriminant ability can thus be obtained. On all data sets, the proposed method achieved a maximum improvement over GCE-L, FS, LS, and LRS of 32.31% (“gesture2” for a label rate α = 1%), 32.80% (“gesture2” for a label rate α = 1%), 14.47% (“vehicle” for a label rate α = 30%), and 2.06% ("spam" for a label rate α = 30%), respectively, when the KNN classifier was used, and a maximum improvement over GCE-L, FS, LS, and LRS by 22.30% ("krvskp" for a label rate α = 1%), 29.10% ("landsat" for a label rate α = 5%), 18.11% ("kdd" for a label rate α = 25%), and 1.48% ("gesture2" for a label rate α = 5%), respectively, when the SVM classifier was used. These results demonstrate the effectiveness of the proposed method.

It is worth mentioning that, on the data sets "cardio," "frog," and "vehicle," the proposed method outperforms Raw. This is probably a consequence of attribute reduction, and confirms the fact that attribute reduction may alleviate overfitting and improve performance. Additionally, on the data sets "gesture1," "musk," and "vehicle," the proposed method achieved an improvement over GT (i.e. the data set with a label rate α = 100%) of 1.89% (KNN for a label rate α = 30%), 4.05% (SVM for a label rate α = 25%), and 5.41% (KNN for a label rate α = 25%), respectively. These results may be because the proposed method selects more informative and discriminative attributes after the data set is labeled through the labeling strategy for unlabeled data. These findings further demonstrate the potential of the proposed method.

To examine the statistical difference, the proposed method was compared with the supervised GCE-L and FS, the unsupervised LS, and the semi-supervised LRS for different label rates. More specifically, the selected methods were first ranked according to their performance, and then the Friedman test was performed on the ranked values. The variable τ_F in the Friedman test was calculated by:

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}}, \tag{11}$$

$$\tau_{\chi^2} = \frac{12N}{k(k+1)}\left(\sum_{i=1}^{k} r_i^2 - \frac{k(k+1)^2}{4}\right), \tag{12}$$

where k is the number of the selected methods, N is the number of data sets, and r_i is the average rank of the i-th method across different data sets.

The variable τ_F follows the F distribution with degrees of freedom $k-1$ and $(k-1)(N-1)$, and the critical value of $F(4,44)$ is 2.0772 when $\alpha = 0.1,\ k = 5$, and $N = 12$. If the selected methods are equivalent in terms of performance, the value of the Friedman statistic should not be greater than the critical value of $F(4,44)$; otherwise, these five methods differ significantly. Tables 6 and 7 show the rank information of the selected methods for different labeled rates.

In Tables 6 and 7, the values in each cell denote the rank of the selected method on each data set with different label rates, and the average rank (underlined). The last row "Avg." indicates the average rank of each method over different data sets.

It can be seen that, when the KNN and SVM classifiers were used, the average ranks of the methods are $[4.04, 4.41, 2.64, 2.43, 1.49]$ and $[4.17, 3.96, 3.17, 2.2, 1.49]$, respectively. The calculated values of τ_F are 15.5976 and 11.8192, respectively, which are greater than the critical value of $F(4,44)$. Therefore, the hypothesis is rejected, and thus the selected methods are significantly different in terms of performance.

A post hoc test was used to further examine the performance difference between the selected methods. The Nemenyi test was selected, the critical value of which was determined by:

$$CD_\alpha = q_\alpha\sqrt{\frac{k(k+1)}{6N}}, \tag{13}$$

where q_α is the critical value of the Tukey distribution at significance level α. The value of $q_{0.1}$ is 2.459 when $\alpha = 0.1$, and then the critical value of $CD_{0.1}$ is 1.5873 $(k = 5, N = 12)$.

Table 6
The rank information of the selected methods for different label rates (KNN).

Data set	GCE-L	FS	LS	LRS	Ours
cardio	5,5,5,5,5,5,5,5	4,4,4,4,4,4,3,3.86	1,1,1.5,2,2,1,1.5,1.43	3,3,3,3,3,3,4,3.14	2,2,1.5,1,1,2,1.5,1.57
frog	5,4,4,3,3,3,3,3.57	4,5,5,5,5,5,5,4.86	3,3,3,4,4,4,4,3.57	2,2,2,2,2,2,2,2	1,1,1,1,1,1,1,1
gesture1	5,4,4,4,4,4,4,4.14	4,5,5,5,5,5,5,4.86	1,1,1,1,1,1,1,1	2,3,3,3,3,3,3,2.86	3,2,2,2,2,2,2,2.14
gesture2	4,4,4,4,4,4,4,4	5,5,5,5,5,5,5,5	1,1,1,1,1,1,1,1	3,3,3,3,3,3,3,3	2,2,2,2,2,2,2,2
kdd	5,5,4,4,4,4,4,4.29	4,4,3,3,3,1,1,2.71	3,3,5,5,5,5,5,4.43	2,2,2,2,2,3,3,2.29	1,1,1,1,1,2,2,1.29
krvskp	5,4,4,4,5,5,5,4.57	4,5,5,5,4,4,4,4.43	1,1,1,1,1,1,1,1	3,3,3,3,3,3,3,3	2,2,2,2,2,2,2,2
landsat	4,4,4,4,4,4,4,4	5,5,5,5,5,5,5,5	3,3,3,3,3,3,3,3	2,2,2,2,2,2,2,2	1,1,1,1,1,1,1,1
libras	2,5,5,4,4,4,3,3.86	3,4,4,5,5,5,5,4.43	1,3,3,3,3,3,4,2.86	4,2,2,2,2,2,2,2.29	5,1,1,1,1,1,1,1.57
musk	4,4,4,4,4,4,4,4	5,5,5,5,5,5,5,5	3,3,3,3,3,3,3,3	2,2,2,2,2,2,2,2	1,1,1,1,1,1,1,1
spam	4,4,4,4,4,4,3,3.86	5,5,5,5,5,5,5,5	1,2,2,2,2,2,2,1.86	3,3,3,3,3,3,4,3.14	2,1,1,1,1,1,1,1.14
vehicle	5,3,3,3,3,3,3,3.29	4,4,4,4,4,4,4,4	3,5,5,5,5,5,5,4.71	1,2,2,2,2,2,2,1.86	2,1,1,1,1,1,1,1.14
wine	2,4,5,4,4,4,4,3.86	5,5,4,3,3,3,3,3.71	1,3,3,5,5,5,5,3.86	3,1,1,1,2,1,2,1.57	4,2,2,2,1,2,1,2
Avg.	4.17,4.17,4.17,3.92,4,4,3.83,4.04	4.33,4.67,4.5,4.5,4.42,4.25,4.17,4.41	1.83,2.42,2.63,2.92,2.92,2.83,2.96,2.64	2.5,2.33,2.33,2.33,2.42,2.42,2.67,2.43	2.17,1.42,1.38,1.33,1.25,1.5,1.38,1.49

Table 7
The rank information of the selected methods for different label rates (SVM)

Data set	GCE-L	FS	LS	LRS	Ours
cardio	5,5,5,5,5,5,5,5	4,4,3,3,3,3,3,3.29	1,1,1,2,2,2,2,1.57	3,3,4,4,4,4,4,3.71	2,2,2,1,1,1,1,1.43
frog	4,4,4,4,4,4,3,3.86	3,3,3,3,3,3,4,3.14	5,5,5,5,5,5,5,5	1,1,1,1,1,1,1,1	2,2,2,2,2,2,2,2
gesture1	5,5,4,4,4,4,4,4.29	4,4,5,5,5,5,5,4.71	1,1,3,3,3,3,3,2.43	3,3,2,2,2,2,1,2.14	2,2,1,1,1,1,2,1.43
gesture2	4,5,4,4,4,4,4,4.14	5,4,5,5,5,5,5,4.86	1,1,1,1,1,1,1,1	3,3,3,3,3,3,3,3	2,2,2,2,2,2,2,2
kdd	5,4,4,4,4,4,4,4.14	4,3,3,2,2,2,2,2.57	3,5,5,5,5,5,5,4.71	1,2,2,3,3,3,3,2.43	2,1,1,1,1,1,1,1.14
krvskp	5,4,4,4,4,4,4,4.14	4,5,5,5,5,5,5,4.86	1,1,1,1,1,1,1,1	2,3,3,3,3,3,3,2.86	3,2,2,2,2,2,2,2.14
landsat	4,4,4,4,4,4,4,4	5,5,5,5,5,5,5,5	3,3,3,3,3,3,3,3	1.5,1.5,1.5,1.5, 1.5,1.5,1.5,1.5	1.5,1.5,1.5,1.5,1.5,1.5,1.5,1.5
libras	3,4.5,3,3,3,3,3,3.21	3,3,5,4,4,4,4,3.86	3,4.5,4,5,5,5,5,4.5	5,2,2,2,2,2,2,2.43	1,1,1,1,1,1,1,1
musk	5,5,5,5,5,5,5,5	4,4,4,4,3,3,3,3.57	3,3,3,3,4,4,4,3.43	1,2,2,2,2,2,2,1.86	2,1,1,1,1,1,1,1.14
spam	5,5,5,5,5,4,3,4.57	4,4,4,3,3,3,4,3.57	3,3,3,4,4,5,5,3.86	1,2,2,2,2,2,2,1.86	2,1,1,1,1,1,1,1.14
vehicle	5,4,3,3,3,3,3,3.43	4,5,5,5,5,4,4,4.57	3,3,4,4,4,5,5,4	1,1,2,2,2,2,2,1.71	2,2,1,1,1,1,1,1.29
wine	2,5,5,5,5,4,4,4.29	5,4,4,3,3,3,3,3.57	1,3,3,4,4,5,5,3.57	3,2,1,2,2,1,2,1.86	4,1,2,1,1,2,1,1.71
Avg.	4.33,4.54,4.17,4.17, 3.83,3.75,3.92,3.96	4.08,4,4.25,3.92, 3.42,3.67,3.67,3.17	2.33,2.79,3,3.33, 2.29,2.21,2.21,2.2	2.13,2.13,2.13,2.29, 1.29,1.38,1.38,1.49	2.13,1.54,1.46,1.29,

Two methods are considered significantly different if the difference of their average rank over different data sets is greater than the critical value of $CD_{0.1}$. By observing Tables 6 and 7, it can be seen that the difference of the average rank of the proposed method and the supervised methods GCE-L and FS over all data sets is greater than $CD_{0.1}$ when the KNN and SVM classifiers were used. Thus, the Nemenyi test demonstrates that the proposed method is significantly better than the supervised methods GCE-L and FS. Moreover, the proposed method is significantly better than the unsupervised method LS when the SVM classifier was used, but the difference is not significant when the KNN classifier was used. Although there is no statistically significant difference, the average rank of the proposed method is consistently better than that of LRS.

5. Conclusions

In many real-world tasks, labeling a large amount of data is exceptionally costly and practically infeasible, and thus available data usually contain a small amount of labeled data but a large amount of unlabeled data. In this paper, we proposed a simple yet effective strategy to generate proxy labels for unlabeled data, incorporating prior knowledge about the entire data set, and considering the class distribution of initially labeled data. To obtain a high-quality reduct of partially labeled data with proxy labels, we integrated information granularity with conditional entropy, and developed a novel granular conditional entropy, which was theoretically proved to be a monotonic attribute reduction measure. Moreover, a heuristic algorithm based on the proposed granular conditional entropy was designed to quickly obtain the optimal reduct of partially labeled data. Experimental results on several benchmark data sets demonstrated that the proposed method is effective for partially labeled data, and is even better than the supervised method on the entire data set with true labels. It should be noted that, to handle numerical attributes, a discretization pre-processing process is involved, and thus an extended method that could directly handle both categorical and numerical attributes should be developed. Another possible direction is to explore an iterative labeling strategy using fuzzy clustering to further improve the quality of proxy labels.

CRediT authorship contribution statement

Can Gao: Conceptualization, Methodology, Software, Data curation, Writing - original draft. **Jie Zhou:** Methodology. **Duoqian Miao:** Methodology, Writing - review & editing. **Xiaodong Yue:** Software, Validation. **Jun Wan:** Software, Visualization.

Declaration of Competing Interest

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

Acknowledgements

The author would like to thank the Editor-in-Chief, editors and anonymous reviewers for their kind help and valuable comments. The work was supported in part by the National Natural Science Foundation of China (Nos. 61806127, 62076164, 62002233), in part by Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515011861), in part by Shenzhen Institute of Artificial Intelligence and Robotics for Society, and in part by the Bureau of Education of Foshan (No. 2019XJZZ05).

Appendix A. Appendix

Proof of Proposition 2.

$$
\begin{aligned}
\mathit{\Delta} GH &= GH(D|P) - GH(D|Q)\\
&= P(X_i)^2 \sum_{k=1}^{|U/D|} P(Y_k|X_i)P(Y_k|X_i) + P(X_j)^2 \sum_{k=1}^{|U/D|} P(Y_k|X_j)P(Y_k|X_j) - P(X_{ij})^2 \sum_{k=1}^{|U/D|} P(Y_k|X_{ij})P(Y_k|X_{ij})\\
&= \tfrac{1}{|U|^2} \sum_{k=1}^{|U/D|} |X_i|^2 \tfrac{|X_i \cap Y_k|}{|X_i|} \log \tfrac{|X_i \cap Y_k|}{|X_i|} + |X_j|^2 \tfrac{|X_j \cap Y_k|}{|X_j|} \log \tfrac{|X_j \cap Y_k|}{|X_j|} - |X_{ij}|^2 \tfrac{|X_{ij} \cap Y_k|}{|X_{ij}|} log \tfrac{|X_{ij} \cap Y_k|}{|X_{ij}|}
\end{aligned}
$$

Let $\theta_i = |X_i \cap Y_k|/|X_i|, \theta_j = |X_j \cap Y_k|/|X_j|$, and $\theta_{ij} = |X_{ij} \cap Y_k|/|X_{ij}|$. We have $\Delta GH = \frac{1}{|U|^2} \sum_{k=1}^{|U/D|} \left(|X_i|^2 \theta_i \log \theta_i + |X_j|^2 \theta_j \log \theta_j - |X_{ij}|^2 \theta_{ij} \log \theta_{ij}\right)$ and $|X_{ij}|\theta_{ij} = |X_i|\theta_i + |X_j|\theta_j$ for any decision Y_k.

Let$f_k = |X_i|^2 \theta_i \log \theta_i + |X_j|^2 \theta_j \log \theta_j - |X_{ij}|^2 \theta_{ij} \log \theta_{ij}$. Then, we have $\Delta GH = \frac{1}{|U|^2} \sum_{k=1}^{|U/D|} f_k$. For any k, we have

$$
\begin{aligned}
f_k &= |X_i|^2 \theta_i \log \theta_i + |X_j|^2 \theta_j \log \theta_j - |X_{ij}|^2 \theta_{ij} \log \theta_{ij}\\
&= |X_i|^2 \theta_i \log \theta_i + |X_j|^2 \theta_j \log \theta_j - |X_{ij}| (|X_i|\theta_i + |X_j|\theta_j) \log \left(\tfrac{|X_i|\theta_i + |X_j|\theta_j}{|X_{ij}|}\right)\\
&\geqslant |X_i|^2 \theta_i \log \theta_i + |X_j|^2 \theta_j \log \theta_j - \left(|X_i|^2 \theta_i + |X_j|^2 \theta_j\right) \log \left(\tfrac{|X_i|\theta_i + |X_j|\theta_j}{|X_{ij}|}\right)\\
&= |X_i|^2 \theta_i \left(\log \theta_i - \log \left(\tfrac{|X_i|\theta_i + |X_j|\theta_j}{|X_{ij}|}\right)\right) + |X_j|^2 \theta_j \left(log \theta_j - log \left(\tfrac{|X_i|\theta_i + |X_j|\theta_j}{|X_{ij}|}\right)\right)\\
&= |X_i|^2 \theta_i \log \left(\tfrac{|X_i|\theta_i + |X_j|\theta_i}{|X_i|\theta_i + |X_j|\theta_j}\right) + |X_j|^2 \theta_j \log \left(\tfrac{|X_i|\theta_j + |X_j|\theta_j}{|X_i|\theta_i + |X_j|\theta_j}\right).
\end{aligned}
$$

Let the right hand side of the above inequality be f'_k, and $m = |X_i|, n = |X_j|, \mu = |X_i|\theta_i$, $v = |X_j|\theta_j, \lambda = \theta_i/\theta_j$. We have

$$
f'_k(\mu, v, \lambda) = m\mu \log \left(\frac{\mu + \lambda v}{\mu + v}\right) + nv \log \left(\frac{\mu/\lambda + v}{\mu + v}\right).
$$

$f'_k(\mu, v, \lambda)$ is an explicit function of the variables μ, v, and λ. The partial derivative of$f'_k(\mu, v, \lambda)$ with respect to the variable λ is

$$
\begin{aligned}
\tfrac{\partial f'_k(\mu,v,\lambda)}{\partial \lambda} &= m\mu \left(\tfrac{\mu+v}{\mu+\lambda v}\right) v \log 2 - nv \left(\tfrac{\mu+v}{\mu/\lambda+v}\right) \tfrac{\mu}{\lambda^2} \log 2\\
&= \log 2 \left(\mu v (\mu + v) \left(\tfrac{\lambda m - n}{\lambda(\mu + v\lambda)}\right)\right) \begin{cases} < 0, & 0 < \lambda < n/m \\ = 0, & \lambda = n/m. \\ > 0, & \lambda > n/m \end{cases}
\end{aligned}
$$

$f'_k(\mu, v, \lambda)$ is minimized when $\lambda = n/m$. Namely, $|X_i|\theta_i = |X_j|\theta_j$. In this case, $f'_k(\mu, v, \lambda) = 0$ and $\Delta GH = 0$. Thus, $\Delta GH \geqslant 0$ holds for every possible case. The proposition is proved.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.ins.2021.08.067.

References

[1] Y.L. Cheng, Q.H. Zhang, G.Y. Wang, B.Q. Hu, Optimal scale selection and attribute reduction in multi-scale decision tables based on three-way decision, Inf. Sci. 541 (2020) 36–59.

[2] J.H. Dai, Q.H. Hu, J.H. Zhang, H. Hu, N.G. Zheng, Attribute selection for partially labeled categorical data by rough set approach, IEEE Trans. Cybern. 47 (2017) 2460–2471.

[3] C. Gao, Z.H. Lai, J. Zhou, C.R. Zhao, D.Q. Miao, Maximum decision entropy-based attribute reduction in decision-theoretic rough set model, Knowl.-Based Syst. 143 (2018) 179–191.

[4] C. Gao, Z.H. Lai, J. Zhou, J.J. Wen, W.K. Wong, Granular maximum decision entropy-based monotonic uncertainty measure for attribute reduction, Int. J. Approx. Reason. 104 (2019) 9–24.

[5] C. Gao, J. Zhou, D. Miao, J.J. Wen, X.D. Yue, Three-way decision with co-training for partially labeled data, Inf. Sci. 544 (2021) 500–518.

[6] Y.T. Guo, E.C.C. Tsang, W.H. Xu, D.G. Chen, Local logical disjunction double-quantitative rough sets, Inf. Sci. 500 (2019) 87–112.

[7] M. Hu, E.C.C. Tsang, Y.T. Guo, D.G. Chen, W.H. Xu, A novel approach to attribute reduction based on weighted neighborhood rough sets, Knowl. Based Syst. 220 (2021) 106908.

[8] M. Hu, E.C.C. Tsang, Y.T. Guo, W.H. Xu, Fast and robust attribute reduction based on the separability in fuzzy decision systems, IEEE Trans. Cybern. (2021), https://doi.org/10.1109/TCYB.2020.3040803.

[9] X.Y. Jia, W.H. Liao, Z.M. Tang, L. Shang, Minimum cost attribute reduction in decision-theoretic rough set models, Inf. Sci. 219 (2013) 151–167.

[10] F. Jiang, Y.S. Sui, L. Zhou, A relative decision entropy-based feature selection approach, Pattern Recognit. 48 (7) (2015) 2151–2163.

[11] B.Y. Li, J.M. Xiao, X.H. Wang, Feature selection for partially labeled data based on neighborhood granulation measures, IEEE Access 7 (2019) 37238–37250.
[12] J.D. Li, K.W. Cheng, S.H. Wang, F. Morstatter, R.P. Trevino, J.L. Tang, H. Liu, Feature selection: A data perspective, ACM Comput. Surv. 50 (2018) 1–45.
[13] D.C. Liang, W. Cao, Z.S. Xu, M.W. Wang, A novel approach of two-stage three-way co-opetition decision for crowdsourcing task allocation scheme, Inf. Sci. 559 (2021) 191–211.
[14] J.Y. Liang, Z.Z. Shi, The information entropy, rough entropy and knowledge granulation in rough set theory, Int. J. Uncertain. Fuzziness, Knowl.-Based Syst. 12 (2004) 37–46.
[15] P. Lingras, M. Chen, D.Q. Miao, Semi-supervised rough cost/benefit decisions, Fundam. Informaticae 94 (2009) 233–244.
[16] K.Y. Liu, E.C.C. Tsang, J.J. Song, H.L. Yu, X.J. Chen, X.B. Yang, Neighborhood attribute reduction approach to partially labeled data, Granul. Comput. 5 (2020) 239–250.
[17] K.Y. Liu, X.B. Yang, H.L. Yu, J.X. Mi, P.X. Wang, X.J. Chen, Rough set based semi-supervised feature selection via ensemble selector, Knowl.-Based Syst. 165 (2019) 282–296.
[18] X.A. Ma, G.Y. Wang, H. Yu, T.R. Li, Decision region distribution preservation reduction in decision-theoretic rough set model, Inf. Sci. 278 (2014) 614–640.
[19] D.Q. Miao, C. Gao, N. Zhang, Z.F. Zhang, Diverse reduct subspaces based co-training for partially labeled data, Int. J. Approx. Reason. 52 (2011) 1103–1117.
[20] F. Min, F.L. Liu, L.Y. Wen, Z.H. Zhang, Tri-partition cost-sensitive active learning through kNN, Soft Comput. 23 (2019) 1557–1572.
[21] P. Ni, S.Y. Zhao, X.Z. Wang, H. Chen, C.P. Li, PARA: A positive-region based attribute reduction accelerator, Inf. Sci. 503 (2019) 533–550.
[22] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (1982) 341–356.
[23] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishers, Dordrecht, Netherlands, 1991.
[24] Z. Pawlak, S.K.M. Wong, W. Ziarko, Rough sets: Probabilistic versus deterministic approach, Int. J. Man-Mach. Stud. 29 (1988) 81–95.
[25] J. Qian, C.L. Liu, D.Q. Miao, X.D. Yue, Sequential three-way decisions via multi-granularity, Inf. Sci. 507 (2020) 606–629.
[26] Y.H. Qian, J.Y. Liang, Combination entropy and combination granulation in rough set theory, Int. J. Uncertainty, Fuzziness Knowl. Based Syst. 16 (2) (2008) 179–193.
[27] Y.H. Qian, X.Y. Liang, G.P. Lin, Q. Guo, J.Y. Liang, Local multigranulation decision-theoretic rough sets, Int. J. Approx. Reason. 82 (2017) 119–137.
[28] Y.H. Qian, X.Y. Liang, Q. Wang, J.Y. Liang, B. Liu, A. Skowron, Y.Y. Yao, J.M. Ma, C.Y. Dang, Local rough set: A solution to rough data analysis in big data, Int. J. Approx. Reason. 97 (2018) 38–63.
[29] R. Sheikhpour, M.A. Sarram, S. Gharaghani, M.A.Z. Chahooki, A Survey on semi-supervised feature selection methods, Pattern Recognit. 64 (2017) 141–158.
[30] L. Sun, J.C. Xu, Y. Tian, Feature selection using rough entropy-based uncertainty measures in incomplete decision systems, Knowl. Based Syst. 36 (2012) 206–216.
[31] K. Thangavel, A. Pethalakshmi, Dimensionality reduction based on rough set theory: A review, Appl. Soft Comput. 9 (2009) 1–12.
[32] C.Z. Wang, Y. Huang, M.W. Shao, Q.H. Hu, D.G. Chen, Feature selection based on neighborhood self-information, IEEE Trans. Cybern. 50 (2020) 4031–4042.
[33] C.Z. Wang, Y. Huang, W.P. Ding, Z.H. Cao, Attribute reduction with fuzzy rough self-information measures, Inf. Sci. 549 (2021) 68–86.
[34] G.Q. Wang, T.R. Li, P.F. Zhang, Q.Q. Huang, H.M. Chen, Double-local rough sets for efficient data mining, Inf. Sci. 571 (2021) 475–498.
[35] Q. Wang, Y.H. Qian, X.Y. Liang, Q. Guo, J.Y. Liang, Local neighborhood rough set, Knowl. Based Syst. 153 (2018) 53–64.
[36] W.H. Xu, Y.T. Guo, Generalized multigranulation double-quantitative decision-theoretic rough set, Knowl. Based Syst. 105 (2016) 190–205.
[37] W.H. Xu, M.M. Li, X.Z. Wang, Information fusion based on information entropy in fuzzy multi-source incomplete information system, Int. J. Fuzzy Syst. 19 (2017) 1200–1216.
[38] W.H. Xu, W.T. Li, Granular computing approach to two-way learning based on formal concept analysis in fuzzy datasets, IEEE Trans. Cybern. 46 (2016) 366–379.
[39] W.H. Xu, J.H. Yu, A novel approach to information fusion in multi-source datasets: A granular computing viewpoint, Inf. Sci. 378 (2017) 410–423.
[40] J.L. Yang, Y.Y. Yao, A three-way decision based construction of shadowed sets from Atanassov intuitionistic fuzzy sets, Inf. Sci. 577 (2021) 1–21.
[41] J.T. Yao, A.V. Vasilakos, W. Pedrycz, Granular computing: Perspectives and challenges, IEEE Trans. Cybern. 43 (2013) 1977–1989.
[42] Y.Y. Yao, Three-way decisions with probabilistic rough sets, Inf. Sci. 180 (2010) 341–353.
[43] Y.Y. Yao, The superiority of three-way decisions in probabilistic rough set models, Inf. Sci. 181 (2011) 1080–1096.
[44] Y.Y. Yao, Three-way decision and granular computing, Int. J. Approx. Reason. 103 (2018) 107–123.
[45] Y.Y. Yao, Three-way granular computing, rough sets, and formal concept analysis, Int. J. Approx. Reason. 116 (2020) 106–125.
[46] Y.Y. Yao, Tri-level thinking: Models of three-way decision, Int. J. Mach. Learn. Cybern. 11 (2020) 947–959.
[47] X.D. Yue, Y.F. Chen, D.Q. Miao, H. Fujita, Fuzzy neighborhood covering for three-way classification, Inf. Sci. 507 (2020) 795–808.
[48] P.F. Zhang, T.R. Li, G.Q. Wang, C. Luo, H.M. Chen, J.B. Zhang, D.X. Wang, Z. Yu, Multi-source information fusion based on rough set theory: A review, Inf. Fusion 68 (2021) 85–117.
[49] X. Zhang, C.L. Mei, D.G. Chen, J.H. Li, Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy, Pattern Recognit. 56 (2016) 1–15.
[50] Z.H. Zhou, A brief introduction to weakly supervised learning, Natl. Sci. Rev. 5 (2018) 48–57.