Applied Intelligence (2021) 51:3212–3225
https://doi.org/10.1007/s10489-020-02014-6

A co-training method based on entropy and multi-criteria

Jia Lu[1] · Yanlu Gong[2]

Accepted: 9 October 2020 / Published online: 10 November 2020

Abstract
Co-training method is a branch of semi-supervised learning, which improves the performance of classifier through the complementary effect of two views. In co-training algorithm, the selection of unlabeled data often adopts the high confidence degree strategy. Obviously, the higher confidence of data signifies the higher accuracy of prediction. Unfortunately, high confidence selection strategy is not always effective in improving classifier performance. In this paper, a co-training method based on entropy and multi-criteria is proposed. Firstly, the data set is divided into two views with the same amount of information by entropy. Then, the clustering criterion and confidence criterion are adopted to select unlabeled data in view 1 and view 2, respectively. It can solve the problem that high confidence criterion is not always valid. Different choices can better play the complementary role of co-training, thus supplement what the other view does not have. In addition, the role of labeled data is fully considered in multi-criteria in order to select more valuable unlabeled data. Experimental results on several UCI data sets and one artificial data set show the effectiveness of the proposed algorithm.

Keywords Co-training · Entropy · Multi-criteria · Naive Bayes · Multi-view

1 Introduction

Semi-supervised learning (SSL) [1, 2] is one of the branches of machine learning, which uses a small number of labeled data and a large number of unlabeled data to obtain high performance classifier, and lacking enough labeled data is the characteristic of most data sets in the real world.

There are many methods in SSL such as self-training methods [3, 4], graph-based method [5], generative model [6] and co-training methods [7, 8]. In co-training algorithm, two different classifiers on two different views are trained firstly, then unlabeled data are labeled through the complementary effects of two views. The process of standard co-training algorithm is shown as follows: 1) The two different classifiers are trained on two different views by labeled data; 2) The trained classifiers are used to classify unlabeled data; 3) Unlabeled data with high confidence are selected, and they are added to the other view respectively; 4)The classifiers, the labeled data set and the unlabeled data set are updated; 5) Repeat 3) and 4) until the iteration stop condition is satisfied.

Standard co-training algorithm requires two sufficient and redundant views, that is, the attributes can be naturally partitioned into two sets, each set is sufficient to learn a classifier with good performance, and they are both independent of each other [9]. But in practice, such conditions are difficult to be satisfied. Work [10] mentioned that the co-training algorithm had acceptable performance of using the randomly divided feature set when it learned from labeled data and unlabeled data. Work [11] proposed a heuristic division to split single view data sets into two views, and it could make co-training work reliably well. Tri-training [12] used the full view of three random data to label unlabeled data, and this method relaxed the condition that co-training algorithm requires two fully redundant independent views.

In co-training algorithm, the selection of unlabeled data has a great influence on the classifier performance. Zhang and Zhou [13] utilized data editing technique to estimate the confidence degree of unlabeled data, and the finding of Angluin and Laird [14] were adopted on the selected data to avoid generating noise points. Work [15] proposed a semi-supervised learning framework combining clustering and classification, which made use of the semi-supervised fuzzy c-means algorithm to find data spatial structure hidden by unlabeled data. However, Euclidean distance could not reflect the correlation between attributes, and it was difficult to choose

✉ Jia Lu
jia-lun@163.com

[1] College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, People's Republic of China

[2] College of Computer Science, Chongqing University, Chongqing 401331, People's Republic of China

appropriate parameters. Then Gong and Lv [16] used semi-supervised metric based fuzzy clustering (SMUC) to select unlabeled data. This method could select unlabeled data with implication information, but its calculation cost was too high. Thus, a co-training method combined with active learning and density peaks clustering (CTALDP) was proposed [17], which utilized density peak clustering [18] to select unlabeled data instead of semi-supervised clustering. The above methods all select data by high confidence criterion, the reason is that high confidence data are more likely to be predicted correctly [19].

Unfortunately, it is not always able to improve the classification performance with the high confidence unlabeled data [20]. Work [20] not only considered the high confidence of the data, but also added the nearest neighbor criterion in the selection of unlabeled data. Work [21] chose the data which were located at the margin of the current classifier and the confidence level above the preset threshold. Another problem in co-training algorithm is that the information of labeled data is not used in the process of selecting unlabeled data, but labeled data often provide more correct hidden information than unlabeled data, no matter how high confidence of selected unlabeled data.

In order to solve these problems above, a new co-training algorithm based on the entropy and multi-criteria (CTEMC) is proposed. Specifically, the view partitioning adopts entropy to get two views with the same amount of information, the selection of unlabeled data use multi-criteria to solve the problem that high confidence criterion is not always effective. Another advantage of multi-criteria is that it takes into account the difference between the two views of co-training, and the role of difference in co-training has been confirmed by many studies [22, 23]. Moreover, in CTEMC, labeled data are utilized to find more reliable unlabeled data. The major contributions of CTEMC are two-fold:

(1) A novel co-training framework is defined, which uses two different criteria to select unlabeled data in two views and then they are added to the other view. The complementarity of co-training guarantees mutual promotion on two criteria.
(2) In the two selection criteria, the effect of labeled data is fully considered, because the information of labeled data is more accurate than unlabeled data.

The rest of the paper is organized as follows. Section 2 describes related theory including entropy, K-means clustering algorithm and Naive Bayes algorithm. Section 3 introduces view partitioning, the two criteria of the proposed algorithm and presents the algorithm in detail. Experimental results on real data sets and further analysis are given in Section 4. Section 5 concludes this paper and indicates several issues for future work.

2 Related theory

Co-training algorithm was proposed by Blum and Mitchell [24], which was inspired by the task of learning to classify web pages, and they wanted to use a large number of unlabeled data to improve the performance of the classifier. The co-training method is simple and effective, it has been extensively researched. In co-training, the two views of the data set may exist naturally, or they can be obtained by dividing a relatively large attribute set. A common variant of co-training algorithm is to generate different classifiers by using different learning algorithms. The decision tree can divide the data space into several equivalence classes, thus Goldman and Zhou [22] used two different decision tree algorithms to get classifiers. Hady and Schwenker [23] proposed the framework of Co-Training by Committee (CoBC), using different classifiers to learn from each other during the training process, and the difference between the classifiers effectively improves the performance of the collaborative training algorithm. Moreover, tri-training [12] is also one of the co-training variant algorithm. The above methods all want to use the difference to obtain a better-performing classifier.

Many researches on co-training are about the selection of unlabeled data [13, 15–17], however, the information of labeled data are rarely used. Obviously, labeled data can provide more correct information. It is also a waste if the information of labeled data are not used. Therefore, in our proposed method, different selection strategies are used to select unlabeled data on two views. This method makes use of the difference to provide more information to another view. At the same time, labeled data are considered in both clustering criterion and confidence criterion.

The rest of this section introduces some related theories which used in CTEMC, including entropy, K-means clustering algorithm and Naive Bayes algorithm. The entropy is used in view partitioning to estimate the information of each attribute, K-means clustering algorithm is used in clustering criterion to find hidden space structure of data set, and Naive Bayes algorithm is used in confidence criterion to calculate the posterior class probability of unlabeled data.

2.1 Entropy

The concept of Information is abstract, and it is difficult to evaluate in a quantitative form until Shannon [25] proposed mathematical formula named entropy. Generally, if the value of entropy is greater, the uncertainty of the event and the amount of information is greater.That is, the probability of an event occurring is 1 or 0, it means that the event will definitely occur (1) or will not occur (0). In this case, the certainty of the event is very high, the value of entropy is small. On the contrary, if the probability of events occur tends

to be consistent, the uncertainty of the event is high, the value of entropy is larger.

Assume there are some events $S = \{s_1,s_2,...,s_n\}$, where the probability of s_i is $p_i(0 \le p_i \le 1)$, and $p_1 + p_1 + ... + p_n = 1$. So the amount of information of each event s_i can be calculated as formula (1).

$$I(s_i) = -\log_2(p_i) \tag{1}$$

The $I(s_i)$ shows the occurrence of event s_i. It only focus on one event, but in most cases, the $I(S)$ of the entire event set is usually showing more meaningful. Obviously, $I(S)$ can be expressed as follows.

$$I(S) = I(s_1) + I(s_2) + \ldots + I(s_n) \tag{2}$$

The entropy is used to evaluate the average amount of information of the entire event set S, and the best average measure is the expectation of the random event s_i, that is, the calculation formula of entropy is as follows [24].

$$I(S) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{3}$$

2.2 K-means clustering algorithm

K-means clustering algorithm [26] is a common algorithm, which has been widely used [27, 28] because of its simple idea and easy implementation. The result of clustering is that the similarity of data belonging to the same cluster is as high as possible, and the similarity between different clusters is as low as possible. K-means clustering is an unsupervised algorithm and it is often used in data mining. The characteristic of data can be found by k-means clustering, so that the data can be further processed. The division of original K-means algorithm often adopts Euclidean distance. The optimization objective is to minimize the sum of squares of the distance from each datum to each center of the cluster. Define $\boldsymbol{D} = \{x_1,x_2,...,x_m\}$, $\boldsymbol{C} = \{C_1,C_2,...,C_k\}$ is k-mans clustering center, squared error-sum can be expressed as [29]:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - u_i\|_2^2 \tag{4}$$

The calculation formula of u_i is as follows.

$$u_i = \frac{1}{N_i} \sum_{x \in C_i} x \tag{5}$$

Where $\boldsymbol{N}_i$ is the number of examples of cluster i. E describes how closely of examples in the cluster surround the cluster center, the smaller E is, the higher similarity of example in the cluster will be, the effect of clustering is better. Unfortunately, minimizing E is not easy, it is an NP-hard problem, and a greedy strategy is usually used to iteratively find an approximate solution. The iteration process of K-means clustering algorithm is carried out until the iteration stop condition is satisfied: 1) The cluster center is no longer changed or the objective function is minimized; 2) The number of iterations reaches the initial setting.

2.3 Naive Bayes algorithm

Naive Bayes is a classification algorithm based on probability theory [30], and its result represents the uncertainty of all conditions and implicates the degree of confidence in the possibility of different labels. Hence, many researchers used it to estimate the confidence of data [31, 32]. Naive Bayes requires relative independence between the attributes of the data set in theory. Its classification ability is stable and has a small error rate under this condition. This assumption, however, is often invalid in practical applications. So assuming that the attributes are mutually independent to make Naive Bayes algorithm simplified. Although this simplification method reduces the classification effect of the naive Bayes classification algorithm, it greatly simplifies its complexity in actual application scenarios.

Let $\boldsymbol{x} = \{a_1,a_2,...,a_m\}$ denotes a test data, and $a_1,a_2,...,a_m$ are the attributes of $\boldsymbol{x}$, m is the number of the attributes of $\boldsymbol{x}$. Let $\boldsymbol{y} = \{c_1,c_2,...,c_k\}$ denotes the label set, and k denotes the number of the labels. Then the probability $P(a_j|c_i)$, $j = 1,2,...,m$, $i = 1,2,...,k$ are calculated. Assuming that the attributes are independent of each other, the probability of $\boldsymbol{x}$ can be calculated as [33]:

$$P(c_i|x) \propto P(x|c_i)P(c_i) = P(c_i) \prod_{j=1}^{m} P\left(a_j|c_i\right), i = 1, \cdots, k \tag{6}$$

3 A co-training method based on entropy and multi-criteria

The proposed method consists of two parts: the first part is view segmentation, and another part is multi-criteria strategy. In view partitioning, the entropy is used to obtain two views with same amount of information, more description is given in Section 3.1. In multi-criteria strategy, the structure of standard co-training algorithm is changed, and it has two different criteria for selecting unlabeled data in two views, instead of only one criterion in standard co-training algorithm. This change aims to solve the problem that high confidence strategy is not always effective, it can also play the complementary role of co-training through the difference between the two selection criteria, hence more useful unlabeled data can be found. In the multi-criteria, each view has its own strategy to select unlabeled data, and its framework can be described as Fig. 1. View 1 uses clustering criterion (criterion 1) to select unlabeled data, view 2 uses confidence criterion (criterion 2)

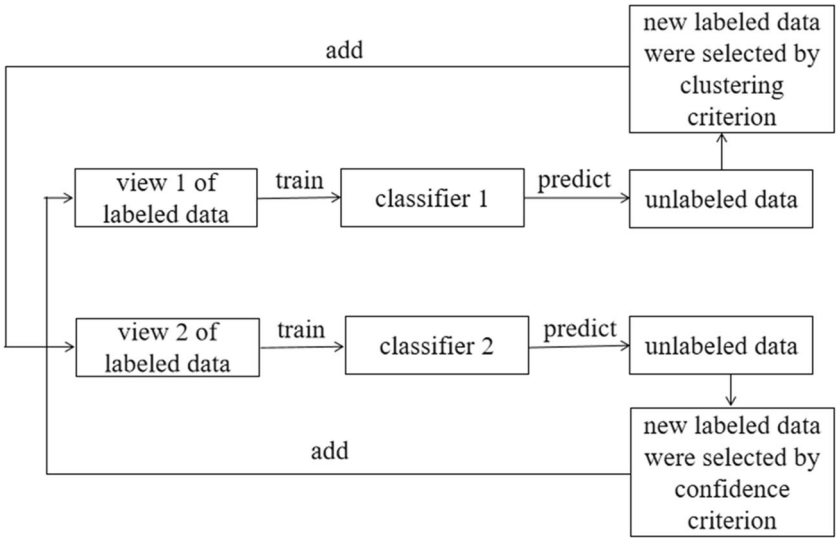

Fig. 1 The framework of multi-criteria co-training

to select unlabeled data. The selected unlabeled data are added to the opposite view respectively to achieve the complementary effect of co-training method.

Labeled data can provide more correct information than unlabeled data, so labeled data are emphasized in both clustering criterion and confidence criterion. Figure 2 is an example of the role of labeled data. Red circle represents labeled data, and blue rhombus represents unlabeled data. In this figure, the confidence of unlabeled data 'b' is 0.988, and the confidence of unlabeled data 'a' is 0.985. From a numerical point of view, the confidence of point 'a' is lower than the confidence of point 'b', but the distance between instance 'b' and labeled data is much higher than instance 'a' to labeled data, then we think unlabeled data 'a' would be a better choice than unlabeled data 'b'.

Specific clustering criterion and confidence criterion are given in 3.2 and 3.3, the whole methods are described in 3.4.

3.1 View partitioning

A simple way to divide a single view into two views is using entropy [34]. The value of entropy is a mathematical representation of the amount of information, thus it can reflect the importance of features.

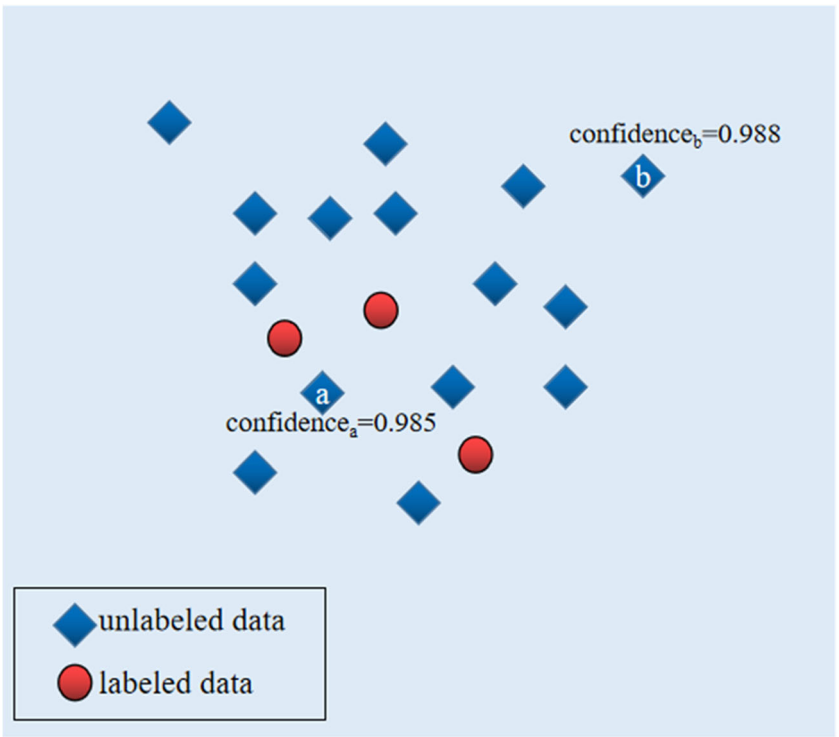

Fig. 2 The example of the role of labeled data

Let $\boldsymbol{D} = \{\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_m\}$ denotes a data set, and $\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_m$ are the features of $\boldsymbol{D}$, m is the number of the attributes of $\boldsymbol{D}$. In order to get two views with the same amount of information, firstly, the entropy of each feature based on the whole data set is calculated, then they are sorted by their entropy. Finally, odd position features are assigned to view 1 and even position features are assigned to view 2. The entropy calculation formula in view partitioning can be written as follows.

$$Info(D_j) = -\sum_{i=1}^{k} p_i \log_2(p_i) \tag{7}$$

Where $\boldsymbol{D}_j$ ($j = 1, \ldots, m$) is the data set which need to calculate the entropy, $\boldsymbol{p}_i$ is the probability that $\boldsymbol{D}_j$ belongs to class i, and k is the number of classes.

3.2 Clustering criterion

K-means clustering algorithm has many advantages, but one of the disadvantage is that the clustering result may be not optimal on account of the randomly selected initial clustering centers.

In the real world, the data in a class consist of many subsets [35]. N. Piroonsup et al. [36] proposed that if the data in a class are clustered into different sub-classes, and the clustering results are more consistent with the data distribution. Therefore, the initial clustering center in the proposed clustering criterion is defined as labeled data. Thus it can avoid the bad effort brought by randomly initial clustering centers. Another superiority is fully considered the information of labeled data. The detailed steps for selecting unlabeled data using clustering criterion are shown in Algorithm 1. At Line 2, the process of clustering criterion will stop when cluster centers are not changed anymore.

Algorithm 1: Apply clustering criterion to select unlabeled data

Input: labeled data ($\boldsymbol{L}$), unlabeled data ($\boldsymbol{U}$), the number of selected unlabeled data (n_1)

Output: A data pool $\boldsymbol{R}_l$

1: Let initial clustering center set $\boldsymbol{C}$ be $\boldsymbol{L}$;

2: **repeat** until stop condition is satisfied

3: Calculate the Euclidean distance $\boldsymbol{d}_{ij}$ between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, $\boldsymbol{x}_i \in \boldsymbol{U}$, $\boldsymbol{x}_j \in \boldsymbol{C}$;

4: Assign $\boldsymbol{x}_i$ to the cluster $\boldsymbol{C}_k$ when $\boldsymbol{d}_{ij}$ is minimum;

5: Calculate mean of each cluster to get the new cluster center set $\boldsymbol{C}$;

6: **end repeat**

7: Select the top n_1 $\boldsymbol{x}$ with the closest distance between the $\boldsymbol{U}$ and the $\boldsymbol{C}$;

8: Add $\boldsymbol{x}$ to data pool $\boldsymbol{R}_l$;

9: **Output** $\boldsymbol{R}_l$

3.3 Confidence criterion

Confidence criterion is a common criterion to select the unlabeled data in co-training algorithm. Nevertheless, researchers often ignore the role of labeled data in confidence criterion. The proposed confidence criterion in this paper uses naive Bayes to obtain the posterior class probability of each datum firstly, those data with high posterior probabilities can be seen high confidence data, then the distance between the data with high confidence and labeled data is calculated. Finally, the high confidence unlabeled data which are closer to labeled data are selected. The selected unlabeled data have two characteristics: 1) Compared with other high-confidence data, the distance between the selected data and labeled data is relatively small; 2) High confidence. Algorithm 2 shows the detailed process of confidence criterion. At Line 2, the confidence of unlabeled data is calculated by Naive Bayes algorithm.

Algorithm 2: Apply confidence criterion to select unlabeled data

Input: labeled data ($\boldsymbol{L}$), unlabeled data ($\boldsymbol{U}$), a learning algorithm l, the number of selected unlabeled data (n_2)

Output: A data pool $\boldsymbol{R}_2$

1: Use l and $\boldsymbol{L}$ to create classifier h;

2: Use h to predict $\boldsymbol{U}$;

3: Select unlabeled data with high confidence and add them to data pool $\boldsymbol{R}_0$;

4: Select the top n_2 $\boldsymbol{x}$ with the closest distance between the $\boldsymbol{U}$ and the $\boldsymbol{R}_0$;

5: Add $\boldsymbol{x}$ to data pool $\boldsymbol{R}_2$;

6: **Output** $\boldsymbol{R}_2$

Table 1 The description of UCI data sets

ID	Data sets	Size	Attribute	class	Abbreviation
1	Pima Indians	768	6	2	PIMA
2	Ecoli	336	8	4	ECOLI
3	abalone	4177	8	3	ABAL
4	Connectionist Bench	208	60	2	CB
5	ionosphere	351	34	2	IONO
6	Banknote authentication	1372	4	2	BA
7	Iris	150	4	3	IRIS
8	seeds	210	7	3	SEED
9	Wine	178	13	3	WINE
10	Breast Cancer Wisconsin (Original)	699	10	2	BCW
11	hcv	615	14	4	HCV
12	Early stage diabetes risk prediction	520	17	2	EDIAB
13	Car Evaluation	1728	6	4	CARE
14	Statlog (Heart)	270	13	2	HSTA
15	Thyroid	215	5	3	TYRD
16	artificial data set	42	2	3	

Table 2 The comparison algorithms and parameters

Algorithm	Abbreviation	Parameters
K-nearest neighbor	KNN	Number of neighbors $K = 3$
Naive Bayes	Naive Bayes	No parameters specified
Standard Co-training algorithm	SCT	The number of selected unlabeled data n = size×10%
Co-training method based on SMUC [16]	CTSMUC	
Co-training combined semi-supervised fuzzy c-means [15]	CTSFCM	
Co-training method combined active learning and density peaks clustering [17]	CTALDP	Threshold of variance $\varepsilon = 0.01$, $d_c = 2$
CTEMC	CTEMC	Threshold of variance $\varepsilon = 0.01$, Euclidean distance, $n_1 = n_2$ = size×10%

3.4 CTEMC

Before the iteration process of CTEMC, view partitioning by entropy is firstly carried out. Secondly, unlabeled data with high ambiguity are selected and they are added into the labeled data set by active learning, this step is the prior work in [17]. Thirdly, unlabeled data are selected according to criterion 1 and they are added into view 2. Similarly, criterion 2 are used to select unlabeled data, then they are added into view 1. Finally, the same data with different labels are found and they are relabeled by the weighted K-nearest neighbor algorithm. The progress will stop when all unlabeled data have been labeled. Algorithm 3 describes the CTEMC in detail.

Algorithm 3: CTEMC

Input: labeled data ($\boldsymbol{L}$), unlabeled data ($\boldsymbol{U}$), a learning algorithm l, the threshold ε

Output: classifier h_1 and h_2

1: Use l and $\boldsymbol{L}$ to create classifier h;

2: Use h to classify $\boldsymbol{U}$, and get the membership degree of each class $\boldsymbol{u}_{ij}$, i=1,...,n (n is the number of unlabeled data $\boldsymbol{U}$), j=1,...,k (k is the number of classes);

3: Calculate variance of the $\boldsymbol{u}_i$ ($\boldsymbol{u}_i$=$\{\boldsymbol{u}_{i1},\boldsymbol{u}_{i2},...,\boldsymbol{u}_{ik}\}$) and get data set $\boldsymbol{var}_i$;

4: Select those unlabeled data which satisfy $\boldsymbol{var}_i$<ε, and add them to $\boldsymbol{U}'$;

5: Label the data of $\boldsymbol{U}'$ and get their class label $\boldsymbol{l}'$;

6: Add $\boldsymbol{U}'$ and $\boldsymbol{l}'$ to $\boldsymbol{L}$, remove $\boldsymbol{U}'$ from $\boldsymbol{U}$;

7: Calculate the entropy of each attribute in L and rank it to get D' ;

8: $\boldsymbol{L}_1 = \left\lfloor \boldsymbol{L}_{D_1}, \boldsymbol{L}_{D_3}, \cdots, \boldsymbol{L}_{D'_{maxodd}} \right\rfloor$, $\boldsymbol{L}_2 = \left\lfloor \boldsymbol{L}_{D_2}, \boldsymbol{L}_{D_4}, \cdots, \boldsymbol{L}_{D'_{maxeven}} \right\rfloor$;

9: **Repeat** until the unlabeled data set is empty

10: Use l and $\boldsymbol{L}_1$ to create classifier h_1, use l and $\boldsymbol{L}_2$ to create classifier h_2;

11: Use clustering criterion to get data pool $\boldsymbol{R}_1$, use confidence criterion to get data pool $\boldsymbol{R}_2$;

12: Use h_1 to classify $\boldsymbol{R}_1$ and get class label $\boldsymbol{l}_1$, use h_2 to classify $\boldsymbol{R}_2$and get class label $\boldsymbol{l}_2$;

13: Select the same data with inconsistent labels in $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$, then use weighted KNN to reclassify them, update $\boldsymbol{l}_1$ and $\boldsymbol{l}_2$;

14: Add $\boldsymbol{R}_1$ and $\boldsymbol{l}_1$ to $\boldsymbol{L}_1$, add $\boldsymbol{R}_2$ and $\boldsymbol{l}_2$ to $\boldsymbol{L}_2$, remove $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ from $\boldsymbol{U}$;

15: **end repeat**

At Lines 1–5, the unlabeled data with high ambiguity are defined and selected. The data set is divided into two parts by entropy at Lines 7–8. At Line 11, $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ are selected by clustering criterion and confidence criterion, they are described in Algorithm 1 and Algorithm 2.

4 Experiments

In order to illustrate the effectiveness of the proposed algorithm, 10 UCI data sets and one artificial data set are selected. Those data sets are different in size, the number of attributes

Table 3 Accuracy of different view partitioning methods when the proportion of labeled data is 10%

Dataset	entropy	view complementarity	random view
PIMA	**66.67**	64.47	65.29
ECOLI	**69.54**	60.35	65.13
ABAL	**51.28**	50.66	50.97
CB	56.55	54.84	**58.90**
IONO	**72.94**	71.51	72.15
BA	64.47	63.30	**65.64**
IRIS	**83.67**	82.33	75.00
SEED	**86.43**	84.52	85.71
WINE	**84.94**	80.97	80.40
BCW	93.21	**93.56**	91.20
HCV	**88.52**	87.61	88.33
EDIAB	**70.87**	67.88	68.27
CARE	**69.45**	64.11	66.92
HSTA	72.04	68.70	**73.33**
TYRD	**82.38**	71.57	80.51
average accuracy	**74.20**	71.09	72.52

(Attribute) and the number of classes(Class). The further description is shown in Table 1. The base classifiers, comparison algorithms and their parameters involved in the experiment are given in Table 2. Naive Bayes and K-nearest neighbor algorithms are selected for analyzing the effect of the algorithms under different base classifiers. More descriptions are shown as follows.

(1) K-nearest neighbor(KNN) algorithm is based on instance, the class label of the unlabeled data can be determined by the class label of the K labeled data closest to this unlabeled data. The KNN algorithm does not have a specific target equation, usually, it uses the distance between the data to measure its similarity.
(2) Naive Bayes algorithm is based on Bayes' theorem and has a solid mathematical foundation. The probabilities that the example belongs to each category are calculated first, and then the category with the largest probability is selected as the class of the example. The naive Bayes algorithm can obtain the class of the example and the probability that the example belongs to each category, so it can not only be used for classification, but also can be used to measure the confidence of the data.

At the experimental stage, the 10-fold cross-validation is adopted. The training set is divided into labeled data and unlabeled data by random way, that is, part of the data in training set are selected randomly to be the labeled data, and the label of the rest of data are removed to obtain unlabeled data.

The algorithm is evaluated by accuracy, and the computation of it can be described as formula (8), where $\boldsymbol{T}_{\text{correct}}$ is the

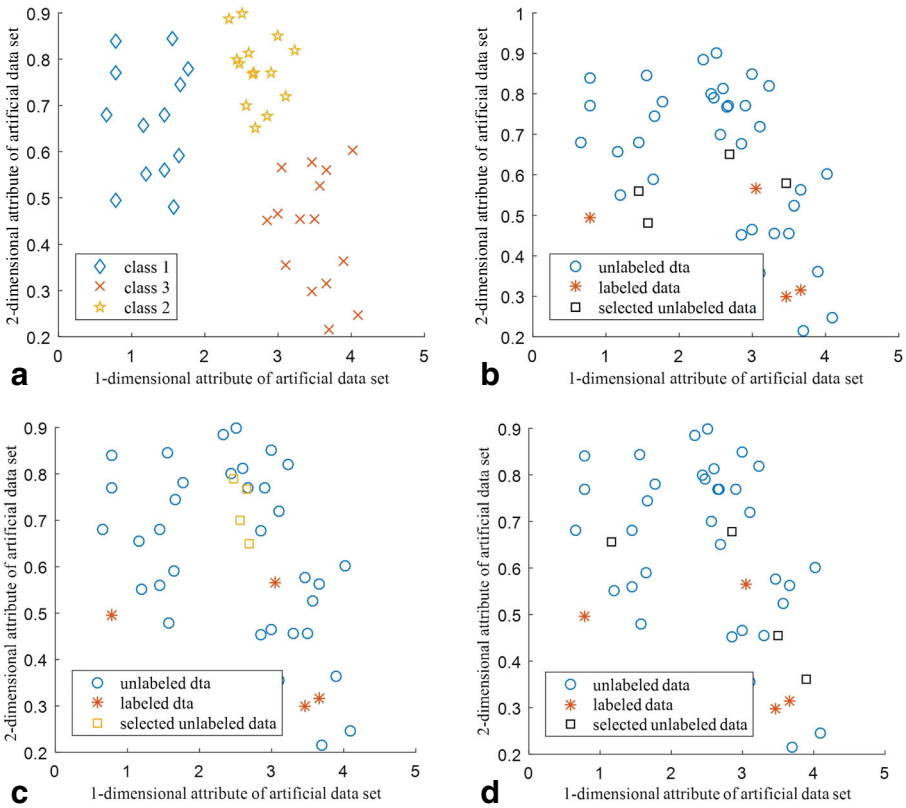

Fig. 3 Schematic diagram of unlabeled data are selected by different methods. **a** Distribution of artificial data set. **b** Unlabeled data are selected by CTSMUC. **c** Unlabeled data are selected by CTALDP. **d** Unlabeled data are selected by propose method

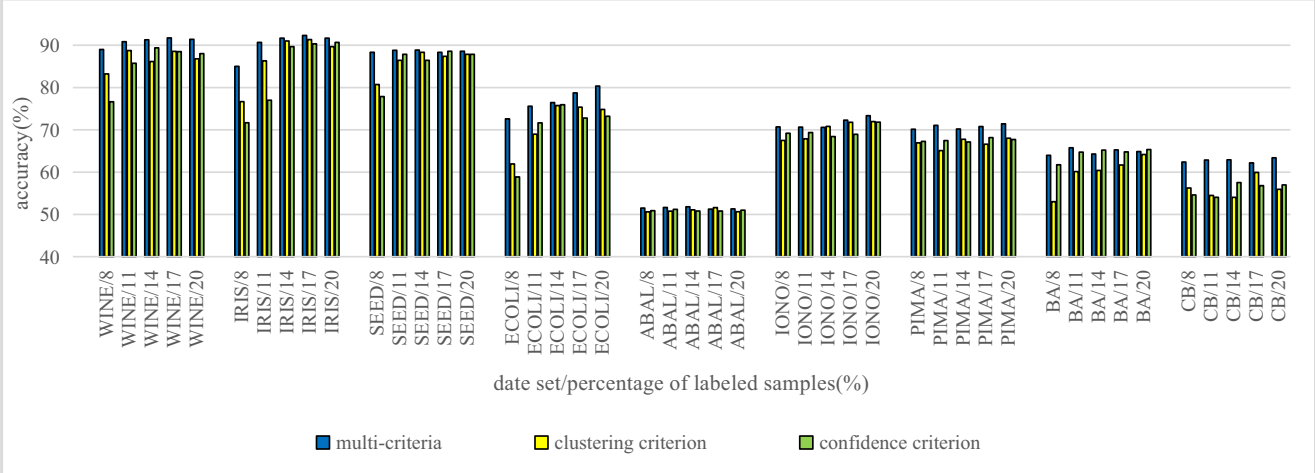

Fig. 4 Result of multi-criteria and single criterion with the base classifier Naive Bayes

number of data which are predicted correctly in testing set and $\boldsymbol{T}$ is the number of testing set.

$$accuracy = \frac{1}{10}\sum_{i=1}^{10}\frac{T_{correct}}{T} \times 100\% \tag{8}$$

4.1 Feature partitioning result

To illustrate the effectiveness of view partitioning based on entropy, a comparative experiment is conducted on 15 UCI data sets. And the another two view partitioning methods are view complementarity and random view. If a data set consists of eight attributes, view complementarity means that view 1 consists of the first four attributes and view 2 is comprised of the remaining attributes. Random view, that is, the data set is divided into two views randomly. The proportion of labeled data is 10% and the experimental results are shown in Table 3.

As shown in Table 3, the accuracy of entropy is higher than view complementary and random view in most cases, the average accuracy of entropy also achieves the highest. It proves the superiority of the entropy to some extent. The performance of view complementary and random view is poor, due to the considerable randomness in feature information.

In the subsequent experiments, CTEMC adopt entropy in view partitioning.

4.2 The result of using the information of labeled data

For analyzing the validity of labeled data, our experiment is carried out on the artificial data set, the percentage of labeled data is 10% and the contrast algorithms are CTSMUC and CTALDP. Theses different methods are utilized to choose unlabeled data which need to be labeled. The results are shown in Fig. 3. Figure 3a is the distribution of artificial data set, it

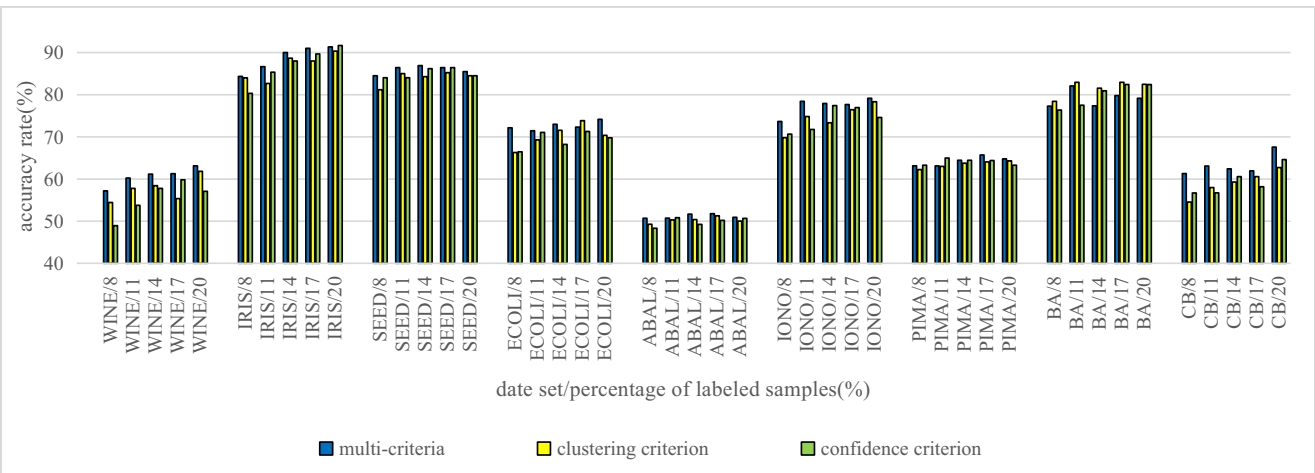

Fig. 5 Result of multi-criteria and single criterion with the base classifier KNN

Table 4 Accuracy of 5 algorithms with the base classifier Naive Bayes when the proportion of labeled data is 8%

Dataset	SCT	CTSMUC	CTALDP	CTSFCM	CTEMC
PIMA	66.93 + (3.20)	67.38 + (2.75)	69.99 + (0.14)	66.67 + (3.46)	**70.13** + (0.00)
ECOLI	64.8 + (7.81)	70.59 + (2.02)	71.74 + (0.87)	61.09 + (11.52)	**72.61** + (0.00)
ABAL	50.79 + (0.70)	51.09 + (0.40)	51.24 + (0.25)	50.97 + (0.52)	**51.49** + (0.00)
CB	53.87 + (8.51)	61.96 + (0.42)	58.31 + (4.07)	56.65 + (5.73)	**62.38** + (0.00)
IONO	70.90-(0.21)	**71.06**-(0.37)	70.12 + (0.57)	70.92-(0.23)	70.69 + (0.00)
BA	63.45 + (0.54)	61.15 + (2.84)	57.70 + (6.29)	62.90 + (1.09)	**63.99** + (0.00)
IRIS	79.33 + (5.67)	77.67 + (7.33)	**89.47**-(4.47)	87.00-(2.00)	85.00 + (0.00)
SEED	84.05 + (4.28)	82.62 + (5.71)	87.62 + (0.71)	85.95 + (2.38)	**88.33** + (0.00)
WINE	73.11 + (15.89)	81.60 + (7.40)	88.89 + (0.11)	76.75 + (12.25)	**89.00** + (0.00)
BCW	92.99 + (0.43)	**93.50**-(0.08)	93.14 + (0.28)	92.49 + (0.93)	93.42 + (0.00)
HCV	80.57 + (7.60)	84.56 + (3.61)	86.89 + (1.28)	80.11 + (8.06)	**88.17** + (0.00)
EDIAB	66.65 + (5.27)	64.71 + (7.21)	67.63 + (4.29)	68.69 + (3.23)	**71.92** + (0.00)
CARE	65.51 + (4.72)	63.74 + (6.49)	64.71 + (5.52)	63.54 + (6.69)	**70.23** + (0.00)
HSTA	65.74 + (6.48)	67.22 + (5.00)	64.63 + (7.59)	66.30 + (5.92)	**72.22** + (0.00)
TYRD	72.42 + (7.58)	73.14 + (6.86)	74.47 + (5.53)	63.30 + (16.7)	**80.00** + (0.00)
average accuracy	70.07	71.47	73.10	70.22	**75.31**

contains 2 attributes and three classes of data, the '◇' represents data of class 1, the ' ' are data of class 2, and the '×' represents data of class 3. Figure 3b, c and d are schematic diagram of unlabeled data are selected by CTSMUC, CTALDP and CTEMC, respectively. The '○', '*' and '□' represent unlabeled data, labeled data and selected unlabeled data.

From Fig. 3, the selected unlabeled data by SMUC are possibly the edge points when the data sets have not obvious boundary line between clusters, and the selected unlabeled data by CTALDP will be clustered in the part with high density.

Figure 3d shows the result of the information of labeled data. As shown in Fig. 3d, if a cluster contains more labeled data than other clusters, the method of labeled data will select more unlabeled data in this cluster, and these data also conform to the distribution of data set.

4.3 The complementarity of co-training algorithm

For the sake of demonstrating that multi-criteria can make use of the complementarity of co-training, we compare multi-criteria with confidence criterion and clustering criterion in

Table 5 Accuracy of 5 algorithms with the base classifier KNN when the proportion of labeled data is 8%

Dataset	SCT	CTSMUC	CTALDP	CTSFCM	CTEMC
PIMA	62.37 + (0.78)	62.50 + (0.65)	63.27-(0.12)	**63.61**-(0.46)	63.15 + (0.00)
ECOLI	64.29 + (7.87)	70.34 + (1.82)	66.86 + (5.30)	66.58 + (5.58)	**72.16** + (0.00)
ABAL	48.11 + (2.60)	49.50 + (1.21)	47.08 + (3.63)	50.19 + (0.52)	**50.71** + (0.00)
CB	58.41 + (2.90)	56.78 + (4.53)	57.73 + (3.58)	57.18 + (4.13)	**61.31** + (0.00)
IONO	72.07 + (1.59)	72.54 + (1.12)	69.08 + (4.58)	69.30 + (4.36)	**73.66** + (0.00)
BA	73.29 + (4.00)	72.27 + (5.02)	**79.54**-(2.25)	74.42 + (2.87)	77.29 + (0.00)
IRIS	74.00 + (10.33)	80.33 + (4.00)	76.00 + (8.33)	82.00 + (2.33)	**84.33** + (0.00)
SEED	75.00 + (9.52)	81.51 + (3.01)	82.38 + (2.14)	79.29 + (5.23)	**84.52** + (0.00)
WINE	52.93 + (4.28)	54.63 + (2.58)	55.02 + (2.19)	54.76 + (2.45)	**57.21** + (0.00)
BCW	93.35-(0.36)	**94.85**-(1.86)	94.28-(1.29)	93.99-(1.00)	92.99 + (0.00)
HCV	87.59 + (0.77)	88.00 + (0.33)	87.10 + (1.23)	**88.33** + (0.00)	**88.33** + (0.00)
EDIAB	71.32 + (4.45)	71.35 + (4.42)	74.97 + (0.80)	73.98 + (1.79)	**75.77** + (0.00)
CARE	61.41 + (4.60)	59.28 + (6.73)	65.04 + (0.97)	63.98 + (2.03)	**66.01** + (0.00)
HSTA	54.81 + (2.41)	50.19 + (7.03)	53.52 + (3.70)	52.81 + (4.41)	**57.22** + (0.00)
TYRD	75.35 + (2.14)	74.40 + (3.09)	75.17 + (2.32)	74.72 + (2.77)	**77.49** + (0.00)
average accuracy	68.29	69.23	69.80	69.68	**72.14**

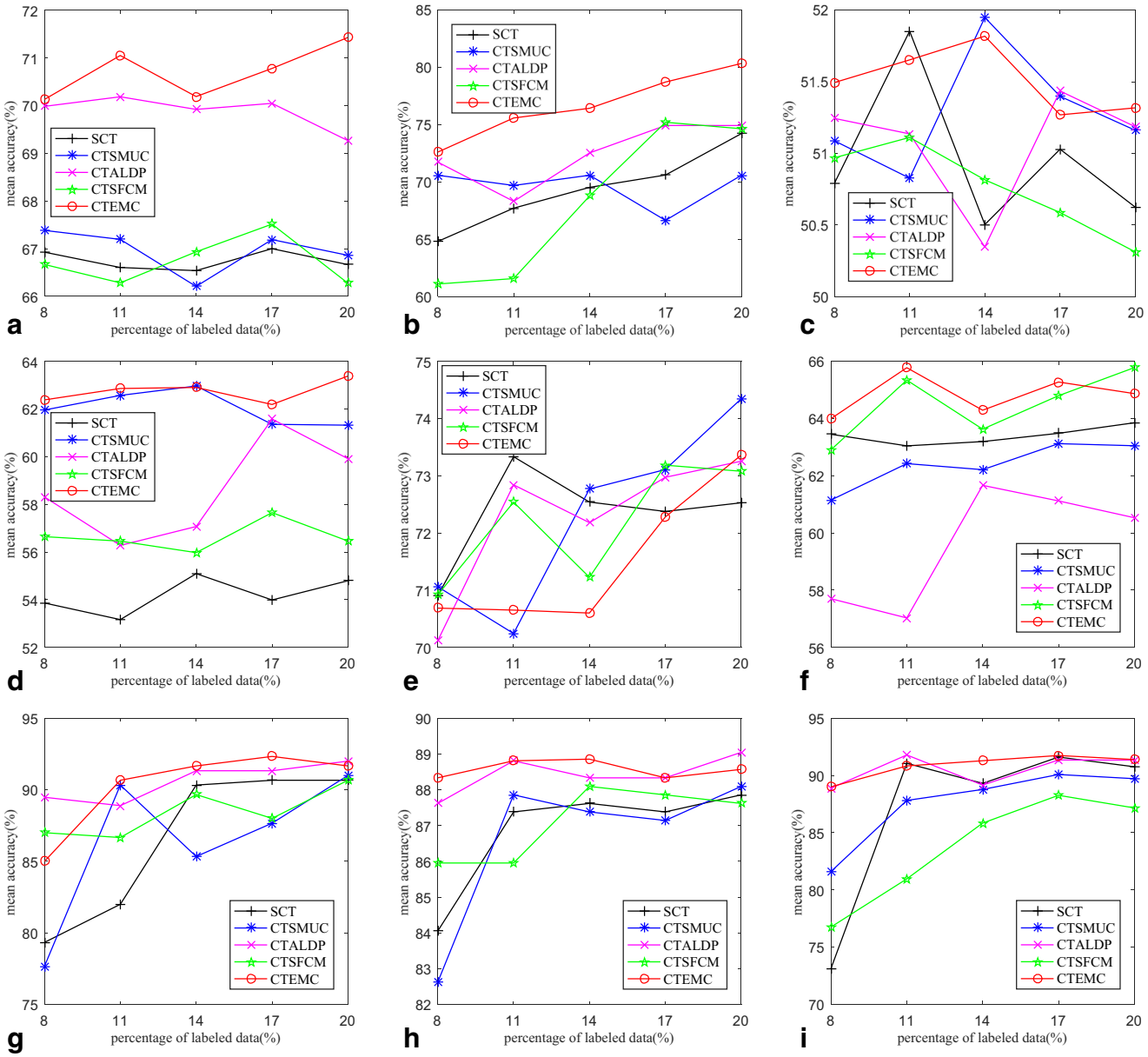

Fig. 6 Relationship chart of classification accuracy and labeled data ratio of 5 algorithms. **a** PIMA. **b** ECOLI. **c** ABAL. **d** CB. **e** IONO. **f** BA. **g** IRIS. **h** SEED. **i** WINE

this section. Figure 4 shows the results with base classifier Naive Bayes, and results with base classifier KNN are shown in Fig. 5.

As shown in Figs. 4 and 5, in most cases, the accuracy of multi-criteria is higher than single clustering criterion or single confidence criterion, which reflects the effectiveness of multi-criteria for selecting unlabeled data.

Table 6 Time complexity of 5 algorithms

Algorithm	Time complexity
SCT	$O(n^2)$
CTSMUC	$O(n^5/m)$
CTALDP	$O(n^2)$
CTSFCM	$O(n^2)$
CTEMC	$O(n^2)$

4.4 The performance of CTEMC

In order to prove the effectiveness of the proposed algorithm, CTEMC is compared with SCT, CTSFCM, CTSMUC and CTALDP.

Tables 4 and 5 are the experimental results of the 5 algorithms on 15 data sets, Naive Bayes is used as the base classifier of co-training algorithm in Table 4 and KNN is used as the base classifier in Table 5, the percentage of labeled data is 8%.

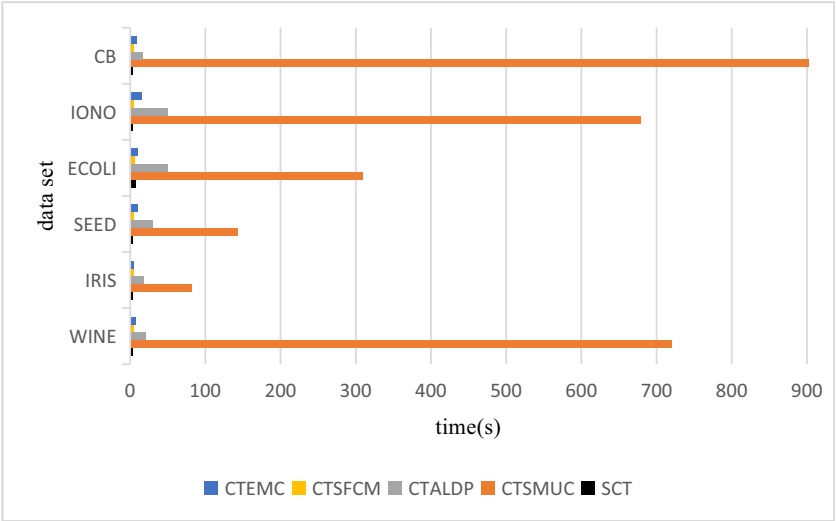

Fig. 7 Result of time consumption with the base classifier Naive Bayes

As shown in Tables 4 and 5, under most circumstances, the accuracy of CTEMC is higher than the contrast algorithms, no matter the base classifier is Naive Bayes or KNN. In detail, 4/5 of the highest accuracy come from CTEMC in Tables 4 and 5. It also has the highest average accuracy on 15 data sets in both Tables 4 and 5, that means the performance of CTEMC is better. In some cases, the accuracy of CTEMC is not the highest in Tables 4 and 5, which may be caused by different characteristics from different base classifiers. On data sets IONO and IRIS, the performance of CTEMC is worse than the contrast algorithms when the base classifier is Naive Bayes, but it is a opposite situation when the base classifier is KNN. Also, on data sets BA and SEED, the accuracy of CTEMC is lower than the contrast algorithms when the base classifier is KNN, it is, however, completely opposite when the base classifier is Naive Bayes. On data set BCW, the algorithm CTSMUC has the highest accuracy, which maybe because the distribution of the BCW is more suitable for SMUC clustering.

In general, the accuracy of CTEMC shows the superiority from Tables 4 and 5, which means the classifier Naive Bayes and KNN can be trained better by CTEMC.

4.5 Effect of the ratio of labeled data

The further discussion about the influence of percentage of labeled data in our algorithm is given in this part. Figure 6 shows the accuracy of 5 algorithms (CTEMC and its contrast algorithms SCT, CTSMUC, CTALDP, CTSFCM) on 9 data sets when the percentage of labeled data are 8%, 11%, 14%, 17% and 20%, respectively.

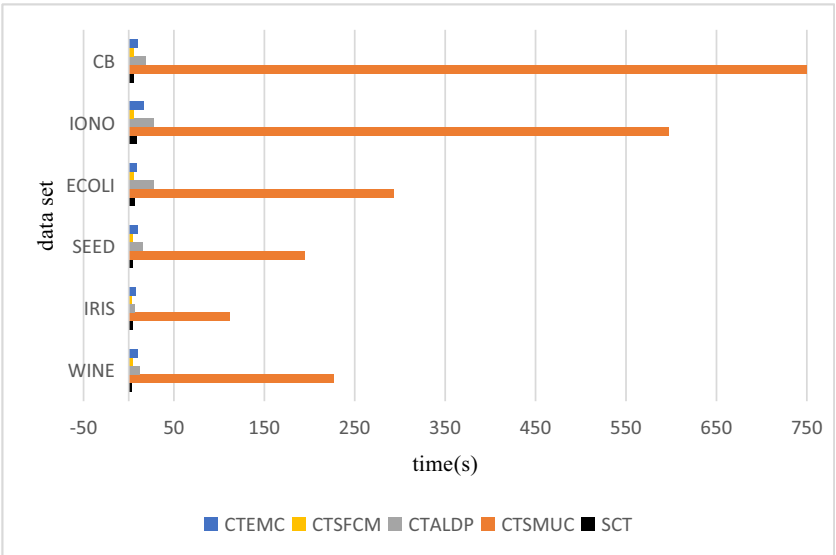

Fig. 8 Result of time consumption with the base classifier KNN

The results are presented in Fig. 6. Observing Fig. 6 we can find that the accuracy of CTEMC is higher than comparison algorithms on PIMA, ECOLI, CB, BA, IRIS, SEED and WINE data sets. It means CTEMC has a better performance in improving accuracy. Compared to contrast algorithm, the accuracy of CTEMC is relative poor on IONO data set, and the reason is that K-means clustering algorithm is not consistent with the distribution of IONO. On ABAL data set, the performance of CTEMC is lower than CTSMUC, but the accuracy of CTEMC is better than SCT, CTALDP, CTSFCM and CTEMC also get a better performance than contrast algorithms when the percentage of labeled data is relatively high.

In general, Fig. 6 shows CTEMC has a better performance under most circumstances.

4.6 Time complexity and time consumption

Finally, The performance of the algorithm is analyzed in terms of time complexity and time consumption. Table 6 is the time complexity and of 5 algorithms, where n is the number of data set and m is the number of selected unlabeled data. The results of time consumption on 6 data sets are shown as Figs. 7 and 8, and the base classifier is Naive Bayes and KNN, respectively. Like Section 4.4 and Section 4.5, the contrast algorithms are SCT, CTSMUC, CTALDP, CTSFCM and CTEMC, the parameters of those algorithms are shown in Table 2, and the percentage of labeled data is 8%.

As can be seen from Table 6, the time complexity of SCT, CTALDP, CTSFCM and CTEMC are equal, and CTSMUC has the highest time complexity. This is because the time-consuming Mahalanobis distance is adopted in SMUC when calculating the distance between data. This phenomenon can also be seen from Figs. 7 and 8, the time consumption of CTEMC (blue column in the histogram), CTSFCM (yellow column), CTALDP (gray column), and SCT (black column) tend to be equal, while the time consumption of CTSMUC (orange column) is much higher than 4 other algorithms. In general, in terms of time consumption and time complexity, the algorithm proposed in this paper does not have a poor performance.

5 Conclusions

In this paper, a new co-training framework based on entropy and multi-criteria is introduced. Firstly, entropy is utilized in view partitioning to get two view sets with the same amount of information. Then, the multi-criteria is put forward to solve the problem that high confidence criterion is not always effective, and it can also take advantage of the complementary of co-training algorithm. Multi-criteria consists of clustering criterion and confidence criterion, and unlabeled data are selected by clustering criterion on view 1 and are added to view 2, unlabeled data are chosen by confidence criterion on the view 2 and are added to the view 1. Besides, the usage of labeled data is strengthened in CTEMC to find more reliable unlabeled data. In detail, the initial clustering center is defined as labeled data in the clustering criterion, and unlabeled data with higher confidence and smaller distance from labeled data are selected in the confidence criterion. In order to verify the effectiveness of the CTEMC, we conduct our experiments on 15 UCI data sets and one artificial data set. The experimental results show that the accuracy of multi-criteria selection of unlabeled data is higher than single criterion, and unlabeled data which are selected by labeled data are also more representative. Compared to SCT, CTSMUC, CTALDP and CTSFCM, CTEMC has superiority in improving accuracy.

In the future, we will discuss the feasibility of other methods in multi-criteria. Additionally, it is very important to enhance the theoretical analysis of our approach. Furthermore, more effective view partitioning methods are also worthy of researching.

Funding This work is supported by Chongqing University Innovation Research Group Founding.

References

1. Gong NZ, Frank M, Mittal P (2017) SybilBelief: a semi-supervised learning approach for structure-based Sybil detection. IEEE Trans Inf Forensics Secur 9(6):976–987
2. Ashfaq RAR, Wang XZ, Huang JZ, Haider A, Yu-Lin H (2017) Fuzziness based semi-supervised learning approach for intrusion detection system. Inf Scienc Int J 378(3):484–497
3. Tanha J, Someren MV, Afsarmanesh H (2017) Semi-supervised self-training for decision tree classifiers. Int J Mach Learn Cybern 8(1):355–370
4. Li J, Zhu Q (2019) Semi-supervised self-training method based on an optimum-path Forest. IEEE Access 7:36388–36399
5. Jiang B, Chen H, Yuan B, Xin Y (2017) Scalable graph-based semi-supervised learning through sparse Bayesian model. IEEE Trans Knowl Data Eng 29(12):2758–2771
6. Meyer SS, Rossiter H, Brookes MJ, Woolrichcet MW, Bestmannd S, Barnesa GR (2017) Using generative models to make probabilistic statements about hippocampal engagement in MEG. NeuroImage 149(2):468–482
7. Zhang X, Song Q, Liu R, Wang W, Jiao L (2017) Modified co-training with spectral and spatial views for Semisupervised Hyperspectral image classification. IEEE J Selected Top Appl Earth Observ Remote Sens 7(6):2044–2055
8. Appice A, Guccione P, Malerba D (2016) A novel spectral-spatial co-training algorithm for the transductive classification of hyperspectral imagery data. Pattern Recogn 63(10):229–245
9. Bin Y, Yang Y, Shen F, Xu X (2016) Combining multi-representation for multimedia event detection using co-training. Neurocomputing 217(23):11–18
10. Zheng Y, Capra L, Wolfson O, Yang H (2014) Urban computing: concepts. Methodol Appl Acm Trans Intell Syst Techno 5(3):1–55

11. Du J, Ling CX, Zhou ZH (2011) When does Cotraining work in real data. IEEE Trans Knowl Data Eng 23(5):788–799
12. Xu C, Tao D, Xu C (2015) Multi-view intact space learning. IEEE Trans Pattern Anal Mach Intell 37(12):1–1
13. Zhang ML, Zhou ZH (2011) COTRADE: confident co-training with data editing. IEEE Trans Syst Man, and Cybern Part B (Cybern) 41(6):1612–1626
14. Angluin D, Laird PD (1988) Learning from Noisy examples. Mach Learn 2(4):343–370
15. Gan H, Sang N, Huang R, Dan Z (2013) Using clustering analysis to improve semi-supervised classification. Neurocomputing 25(3): 290–298
16. GONG YL, LU J (2019) Co-training method combined semi-supervised clustering and weighted K Nearest Neighbor. Comput Eng Appl 55(22):114–118l
17. GONG YL, LU J (2019) Co-training method combined active learning and density peaks clustering. Comput Appl 39(08):2297–2301
18. Rodriguez A, Laio A (2014) Clustering by fast search and find of density peaks. Science 344(6191):1492–1496
19. Sun S (2013) A survey of multi-view machine learning. Neural Comput & Applic 23(7–8):2031–2038
20. Zhang Y, Wen J, Wang X, Jiang Z (2014) Semi-supervised learning combining co-training with active learning. Expert Syst Appl 41(5): 2372–2378
21. Liu ZY, Gao ZB, Li XL (2018) Co-training method based on margin data addition. Chin J Sci Instrum 39(03):45–53
22. Goldman S, Zhou Y (2000) Enhancing supervised learning with unlabeled data, Pcoceedings of the 17th International Conference on Machine Learning San Francisco 327–334
23. Hady MFA, Schwenker F (2008) Co-training by committee: a new semi-supervised learning framework. *IEEE Int Conf Data Min Workshops IEEE*:563–572
24. Blum A, Mitchell T (1998) Combining Labeled and Unlabeled Data with Co-Training, Proceedings of the 11th Annual Conference on Computational Learning Theory
25. Shannon CE (1948) A mathematical theory of communication[J]. Bell Syst Tech J 27(4):379–423
26. Fard MM, Thonet T, Gaussier E (2018) Deep k-means: jointly clustering with k-means and learning representations. Pattern Recogn Lett 138(10):185–192
27. Khanmohammadi S, Adibeig N, Shanehbandy S (2017) An improved overlapping k-means clustering method for medical applications. Expert Syst Appl 67(1):12–18
28. Zhu Q, Pei J, Liu XB, Zhou ZP (2019) Analyzing commercial aircraft fuel consumption during descent: a case study using an improved K-means clustering algorithm. J Clean Prod 223(12): 869–882
29. Liu G, Yang J, Hao Y, Zhang Y (2018) Big data-informed energy efficiency assessment of China industry sectors based on K-means clustering. J Clean Prod 183(9):304–314
30. Abellán J, Castellano JG (2017) Improving the naive Bayes classifier via a quick variable selection method using maximum of entropy. Entropy 19(6):247–264
31. Wang S, Wu L, Jiao L, Liu H (2014) Improve the performance of co-training by committee with refinement of class probability estimations. Neurocomputing 136(8):30–40
32. Dong LY, Sui P, Sun P, Li YL (2016) A new naive bayes classification algorithm based on semi-supervised learning. J Jilin Univ (Eng Edition) 46(3):884–889
33. Feng X, Li S, Yuan C, Zeng P, Sun Y (2018) Prediction of slope stability using naive Bayes classifier. KSCE J Civ Eng 22(3):941–950
34. Nicholson T, Sambridge M, Gudmundsson Ó (2010) On entropy and clustering in earthquake hypocentre distributions. Geophys J R Astron Soc 142(1):37–51
35. Wang Y, Chen S, Zhou ZH (2012) New semi-supervised classification method based on modified cluster assumption. IEEE Trans Neural Netw Learn Syst 23(5):689–702
36. Piroonsup N, Sinthupinyo S (2018) Analysis of training data using clustering to improve semi-supervised self-training. Knowl-Based Syst 143(2):65–80

Jia Lu received her B.S. degree in Computer Science from Chongqing Normal University, Chongqing in 2000. In 2006, she received her M.S. degree in Computer Science and Technology from Chongqing University, China. In 2012, she received her Ph.D. in Applied Mathematics from Inner Mongolia University, Hohhot, Inner Mongolia, China. Her research interests are machine learning and data mining.

Yanlu Gong received her B.S. and M.S. degrees in college of Computer and Information Science from Chongqing Normal University, China, in 2017 and 2020, respectively. She is currently studying for Ph.D. degree in college of Computer Science, Chongqing University. Her current research interests include machine learning, data mining and semi-supervised learning.