This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Cost-Effective Active Semi-Supervised Learning on Multivariate Time Series Data With Crowds

Guoliang He, Bing Li, *Member, IEEE*, Han Wang, and Wenjun Jiang

***Abstract*—Traditional active learning algorithms have several limitations: 1) they cannot obtain satisfactory results on high dimensional datasets, especially for multivariate time series (MTS) data; 2) traditional crowd-based labeling approaches do not consider the swarm intelligence of crowds, which cannot guarantee the confidence of labeling results; and 3) up to now, few works have addressed the issue of trade-off between the labeling accuracy and the labeling cost with crowdsourcing labelers. There is also a lack of research on crowd-based active learning on MTS data. To efficiently address the above issues, we propose a new framework of active semi-supervised learning. First, two criteria are advanced to measure the importance of an unlabeled sample from different perspectives. A dynamic time wrapping (DTW)-based similarity matrix is used to represent the MTS data. Next, to confidently label the most valuable samples with minimum cost, we advance a cost-effective crowd selection model and an adaptive labeler selection approach (ALS) to select the most suitable labelers, which could minimize the total cost of labeling and achieve better classification performance. Experiments on ten datasets show the effectiveness of our proposed methods.**

***Index Terms*—Active learning, crowd labeling, multivariate time series (MTS), sampling.**

I. Introduction

THE SCALE and diversity of the labeled training data is critical to learn a high-quality classification model [1], [2]. However, in real-world applications, there are usually very few labeled samples and a huge number of unlabeled samples. Labeling all the unlabeled samples is time-consuming and expensive, and sometimes it is even impossible. To save the human resources and cost, researchers seek to learn a good model with just a small number of labeled data. Active learning, which labels samples by interactively selecting the most important samples based on certain sampling criteria, therefore has become a hot topic [3]–[5].

In the active learning communities, sampling and annotation are two research topics to be studied. For sampling, its task is to select the most important samples for annotation. It has been widely studied in the past decades and lots of measures have been introduced to rank unlabeled samples, such as uncertainty [6], density [7], information [8], query by committee [9], variance [10], etc. However, it is quite difficult to apply these sampling strategies directly to the multivariate time series (MTS) data. Generally, an MTS sample is a combination of multiple correlated univariate time series, and each univariate time series represents the characteristics of a perspective of the sample. MTS processing unveils plenty of difficulties, among which its high dimensionality is the prominent problem [21]. This specific problem caused by its special data structure also leads to the difficulty in applying the sampling strategies mentioned above directly. On one hand, MTS data is always sparse in spatial distribution, and the length of time series may differ in the same dataset, making it difficult to select valuable unlabeled samples with traditional sampling methods. On the other hand, the chronology and complex structure of MTS data decreases the effectiveness of traditional sampling methods. Up to now, most studies only consider normal data structures and few works have touched sample selection of MTS data.

Meanwhile, taking labeling accuracy and labeling cost into consideration is another research direction for sampling [11]–[13]. Traditional labeling schemes assume that the oracle is absolutely accurate and can provide ground truths to the queried unlabeled samples [14]–[16]. But in practice, labelers usually offer different qualities of labeling for different kinds of samples [17]. To improve the labeling accuracy, crowdsourcing has been studied in recent years, such as the meta-learning ensemble method [18] and NAM/RAM model [19]. There's also a theoretical explanation of the impact of the data collection process on the accuracy of a crowdsourcing system provided by Edoardo *et al.* [20]. However, these studies cannot guarantee that the confidence of annotations reaches a certain level of accuracy. Besides, cost control is another important issue, it is challenging to label an unlabeled sample from crowds with certain accuracy and a lower cost.

Despite the good performance of many state-of-the-art active learning algorithms, there still exists several limitations.

1) They cannot obtain satisfactory results on high dimensional datasets, especially for MTS data.
2) Traditional crowd-based labeling approaches did not consider the swarm intelligence of crowds, therefore cannot guarantee the confidence of labeling results. Since labelers have different accuracy in different kinds

Manuscript received January 18, 2020; revised July 23, 2020; accepted August 15, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0101100; and in part by the National Natural Science Foundation of China under Grant 61876136 and Grant 61832014. This article was recommended by Associate Editor Z. Yu. *(Corresponding authors: Guoliang He; Bing Li.)*

The authors are with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: glhe@whu.edu.cn; bingli@whu.edu.cn; wh20039@whu.edu.cn; 2019202110095@whu.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2020.3019531

2168-2216 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Wuhan University. Downloaded on March 18,2021 at 02:14:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

of samples, only one labeler may not be sufficient to ensure that the labeling accuracy reaches a certain level for an unlabeled sample.

3) Few works have addressed the issue of tradeoff between the labeling accuracy and the labeling cost with crowdsourcing labelers. In a nutshell, it is a challenging issue to identify unlabeled samples with a certain level of accuracy and the minimum cost.

In order to address these limitations and to obtain a high-quality training data with the minimum cost, in this article, we first tackle active semi-supervised learning on MTS with crowds. We propose a cost-effective semi-supervised learning with crowds framework (CSLC), which integrates two valid sampling strategies and a cost-sensitive crowd labeling approach. Compared with traditional active learning methods, CSLC has following properties.

1) In terms of characteristics of MTS data, two valid measure criteria, including informativeness and representativeness are introduced to evaluate the importance of an unlabeled sample from different perspectives. Based on the combination of both criteria, two sampling strategies are proposed to measure the importance of an unlabeled sample.
2) Based on swarm intelligence of crowds, a cost-sensitive crowd labeling module is set to address the issue of confidently labeling an unlabeled sample with the minimum cost.

Next, an optimization algorithm is advanced to obtain the multiple labelers, who can confidently identify the class label of an unlabeled sample with minimum cost. Finally, we evaluate the proposed algorithms on 10 MTS datasets. Experiments show that our methods can drastically obtain a high-quality training data with minimum cost. To the best of our knowledge, the proposed cost-sensitive active semi-supervised learning method is the first one to combine these two issues simultaneously.

The contribution of our work is fourfold. First, to measure the importance of an unlabeled sample from different perspectives, based on dynamic time wrapping (DTW) similarity strategy, we advance two criteria to evaluate its informativeness and representativeness, respectively. Second, a bi-objective-based sample selection model is built to select the most valuable unlabeled sample from a huge number of unlabeled data. Moreover, two selection methods are introduced to select the most valuable sample from unlabeled data for annotation. Third, to confidently label the most valuable samples with minimum cost, we first tackle the issue of the combination of improving the accuracy of labeling and reducing the labeling cost by swarm intelligence, and design a cost-effective crowd selection model. Fourth, based on the proposed crowd selection model, an adaptive labeler selection approach (ALS) is proposed to accurately label unlabeled samples with the lowest cost.

The rest of this article is organized as follows. Section II provides the related work. Section III presents an informativeness and representativeness-based sampling method, and a cost-effective crowd labeling (CCL) approach is proposed in Section IV. Section V introduces an active semi-supervised learning framework. Section VI is about the experiments and Section VII concludes our work and points out future work.

II. Related Work

Active learning focus on selecting valuable samples from vast unlabeled data, saving energy from labeling unimportant data while guaranteeing the model's performance. Recently, works focused on active learning can be divided into two categories.

The first category focuses on how to select the most valuable samples from the unlabeled data. In the last decades, many sample selection strategies have been advanced [6]–[10]. Among them, uncertainty-based sampling is a popular method, aiming to select the most uncertain samples that typically lie to the decision boundary. The assumption is that samples which are harder to identify are more helpful to further enhance the classification performance by updating the classifier. For instance, Lughofer and Pratama [6] advanced two uncertainty-based sampling selection criteria in combination with evolving generalized Takagi–Sugeno (TS) fuzzy models. However, the disadvantage of uncertainty-based sampling is that some outliers can be selected. To handle this issue, density-based sampling was studied to select typical samples for annotation from highly dense regions in the data space. For instance, to handle dynamic data streams, Mohamad *et al.* [22] introduced a density-based criterion, which curbs the sampling bias by weighting the samples to reflect on the true underlying distribution.

To further improve the quality of labeled training data, some works have studied to enhance the diversity of the training data, generally including two typical means. One is clustering methods [8], [23]–[27], and the other is combinations of multiple evaluation criteria [7], [28], [29]. For instance, to reflect the diversity of the bag in multiple-instance active learning, Wang *et al.* [8] proposed two diversity criteria, including clustering-based diversity and fuzzy rough set-based diversity. Meanwhile, some works introduced hybrid strategies to improve the sampling quality. He *et al.* [7] proposed a sampling strategy by ranking the informativeness of unlabeled samples based on its uncertainty and its local data density. To find a general way to choose the most suitable samples, Du *et al.* [30] introduced a systematic and direct way to measure and combine informativeness and representativeness.

The second category of active learning research is labeling strategies. Crowdsourcing systems is a low-cost way to collect labels, but labels obtained are often imperfect. Quality control is a critical problem in crowdsourced data management which includes worker modeling and answer aggregation [31]. Worker modeling is to characterize a worker's quality. Considering multilabel crowdsourcing issue, Li *et al.* [19] proposed two approaches NAM/RAM modeling the crowds' expertise and label correlations from different perspectives. Liu and Liu [32] designed an efficient online algorithm LS_OL using a simple majority voting rule that can differentiate high and low-quality labelers over time. Answer aggregation is another important step which infers the truth from multiple labelers and aggregates all the answers. For instance,

Authorized licensed use limited to: Wuhan University. Downloaded on March 18,2021 at 02:14:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Zhang *et al.* [18] proposed a novel meta-learning ensemble solution for learning from crowds. Instead of inferring true labels of training instances, they could preserve the useful information for learning as much as possible. Based on semi-supervised learning, Atarashi *et al.* advanced a novel generative model of the labeling process in crowdsourcing. It made full use of unlabeled data effectively by introducing latent features and data distribution [33].

Cost control is another issue and there exists a trade-off between quality and cost. Crowdsourcing can collect labels with low labeling costs, which is compatible with the goal of active learning [34], so studies have been done these years to combine crowdsourcing with active learning. To address active learning with imbalanced multiple noisy labeling, Zhang *et al.* [35] proposed a novel active learning framework combining label integration and instance selection into a single method. Hsu and Lin [36] connected active learning with multiarmed bandit problem, letting machines adaptively learn from the performance of the given set of strategies on a particular data set.

Although these crowdsourcing methods could improve the accuracy of labeling, their confidence is not guaranteed because labelers have different accuracy in different kinds of samples. And up to now, most works focus on the cost of sampling [37]–[39], the cost of crowdsourcing is seldom touched. In [17], to improve the labeling accuracy with a low cost, Huang *et al.* proposed a novel active criterion to evaluate the cost-effectiveness of instance-labeler pairs, and selected the labeler who can provide an accurate label for the instance with a relative low cost. To reduce labeling cost, Fu *et al.* [40] made use of nonexpert labelers to carry out the labeling task without explicitly telling the class label of a queried sample.

TABLE I
SYMBOL TABLE

Notation	Definition
D	A training dataset with a few labeled examples and a huge number of unlabeled examples
$\mathcal{L}$	A subset of all labeled samples from the training data
$\mathcal{U}$	A subset of all unlabeled samples from the training data
U_{p}	The nearest positive sample of U
U_{n}	The nearest negative sample of U
U	An unlabeled multivariate time series example U in $\mathcal{U}$
U^*	The selected sample based on the sample selection model
α	The ratio of importance between INFO(X) and REP(X) in LWS
γ	A randomly generated value in the interval of 0 to 1
β	A parameter in control of the proportions of two indicators in IBS
$\mathcal{H}$	Multiple oracles with the scale of m
q_{i}	The labeling accuracy of the example U by the i^{th} labeler in $\mathcal{H}$
δ	The confidence threshold
Θ	A population with N individuals
n	The number of all available labelers
ε	Stable amplitude in the stopping criterion
ϕ	Stable interval in the stopping criterion

III. INFORMATIVENESS AND REPRESENTATIVENESS-BASED SAMPLING

In this section, we first introduce two criteria to evaluate the informativeness and representativeness of an unlabeled sample, respectively. Next, combining both criteria, two efficient selection methods are proposed to select the most important samples from the unlabeled data for annotation. To better present our work, some important notations and definitions in this article are clarified in Table I.

A. Informativeness

Informativeness measures the ability of a sample to help reduce the model's generalization error and uncertainty on classification [28]. To measure the informativeness of an unlabeled sample, we consider its uncertainty and local data density simultaneously. High uncertainty means that the sample lies on the decision boundary, so the current classification models have difficulty identifying its label. And the local data density of an unlabeled sample reflects how much information the learning model can acquire from it. The calculation of informativeness is shown as follows.

Definition 1 (Uncertainty): Given a partial labeled (PL) dataset D, based on the information entropy, we calculate the uncertainty of an unlabeled MTS sample U in $\mathcal{U}$ by

$$\begin{aligned} UCTI(U) &= -(P_U \log P_U + N_U \log N_U) \\ P_U &= \frac{DSim(U, U_p)}{DSim(U, U_p) + DSim(U, U_n)} \\ N_U &= \frac{DSim(U, U_n)}{DSim(U, U_p) + DSim(U, U_n)} \end{aligned} \tag{1}$$

where U_p is the nearest positive sample of U, U_n is the nearest negative sample of U, $DSim(U, U_p)$ is the similarity between U and U_p, and $DSim(U, U_n)$ is the similarity between U and U_n. Among many well-known measurements, such as Euclidean Distance and Mahalanobis Distance, DTW is an elastic method that can compute the alignment between two time series samples. Here, we use DTW-based similarity in this article for its better performance on time series data.

Definition 2 (Reverse k-Nearest Neighbors): Given a PL dataset D, for a MTS S in D, its reverse k-nearest neighbors which is defined as follows:

$$RkNN(S) = \{P \in D | S \in kNN(P)\} \tag{2}$$

where $kNN(P)$ is k-nearest neighbors of sample P.

Definition 3 (Local Spatial Density): Given a PL dataset D, for an unlabeled MTS sample U in $\mathcal{U}$, the local spatial density

Authorized licensed use limited to: Wuhan University. Downloaded on March 18,2021 at 02:14:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

of U is calculated by

$$LSD(U,K) = \frac{1}{K+1} \sum_{X \in \{S\} \cap kNN(U)} |RkNN(X)| \tag{3}$$

where $kNN(U)$ is the k-nearest neighbors of U, $RkNN(X)$ is the reverse k-nearest neighbors of X, and its size is $|RkNN(X)|$.

Based on the uncertainty and the local spatial density of an unlabeled sample U, we can define its informativeness as follows.

Definition 4 (Informativeness): Given a PL dataset D, based on the uncertainty and local data density, for an unlabeled MTS U in $\mathcal{U}$, its informativeness is calculated by

$$INFO(U) = \frac{\min_{X \in kNN(U)} |RkNN(X)| + 1}{|RkNN(U)| + 1} * LSD(U,K) * UCTI(U) \tag{4}$$

where $LSD(U,K)$ is the local spatial density of U, and $UCTI(U)$ is the uncertainty of U.

B. Representativeness

Reprensentativeness measures what extent an instance can represent the whole pattern of the input dataset [28]. Given the dataset D, for an unlabeled sample U in $\mathcal{U}$, its representativeness is represented by the distribution discrepancy between $\mathcal{L}' = \mathcal{L} \cup \{U\}$ and D. Intuitively, the smaller the distribution discrepancy, the more similar the spatial distribution of $\mathcal{L}'$ and D. In our algorithm, we use maximum mean discrepancy (MMD) to evaluate the distribution discrepancy of two datasets [41].

Definition 5 (Maximum Mean Discrepancy): Given two datasets $Q = \{q_1, q_2, \ldots, q_m\}$ with the size of m and $R = \{r_1, r_2, \ldots, r_n\}$ with the size of n. The MMD between Q and R is calculated by

$$MMD(Q,R) = \frac{1}{m(m-1)} \sum_{i \in j}^{m} k(q_i, q_j) + \frac{1}{n(n-1)} \sum_{i \in j}^{n} k(r_i, r_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(q_i, r_j) \tag{5}$$

where $k(x, y)$ is a kernel function determined by the specific data.

Because of the characteristics of MTS data, some well-known kernels, such as Gaussian kernel [42], radial-basis function (RBF) kernel [43], and polynomial kernel [44], are unsuitable for MTS data. So here we introduce a global alignment kernel to measure the MMD. For two MTS samples $X = (X_1, X_2, \ldots, X_m)$ and $Y = (Y_1, Y_2, \ldots, Y_n)$, the global alignment kernel function is calculated by

$$K_{GA}(X,Y) = \sum_{\pi \in A} \prod_{i=1}^{|\pi|} k\big(X_{\pi_1(i)}, Y_{\pi_2(i)}\big)$$
$$k\big(X_{\pi(i)}, Y_{\pi(i)}\big) = e^{-\|X_{\pi(i)} - Y_{\pi(i)}\|^2} \tag{6}$$

where A is the set of all possible alignments between X and Y. An alignment π between X and Y describes the correspondence between each component on the time series X and Y. It is presented as

$$\pi = (\pi_1(\cdot), \pi_2(\cdot)) = \begin{pmatrix} (\pi_1(1), \pi_2(1)) \\ (\pi_1(2), \pi_2(2)) \\ \ldots\ldots \\ (\pi_1(|\pi|), \pi_2(|\pi|)) \end{pmatrix}. \tag{7}$$

Each alignment is subject to some constraints: $\forall 1 \le i \le |\pi| - 1$ and $\forall 1 \le j \le |\pi| - 1$

$$1 = \pi_1(1) \le \cdots \le \pi_1(|\pi|) = m$$
$$1 = \pi_2(1) \le \cdots \le \pi_2(|\pi|) = n$$
$$\pi_1(i+1) \le \pi_1(i) + 1, \pi_2(j+1) \le \pi_2(j) + 1$$
$$(\pi_1(i+1) - \pi_1(i)) + (\pi_2(j+1) - \pi_2(j)) \ge 1.$$

To effectively calculate the global alignment kernel function, we adopt the idea of dynamic programming. For two MTS sample $X = (X_1, X_2, \ldots, X_i, \ldots, X_m)$ and $Y = (Y_1, Y_2, \ldots, Y_j, \ldots, Y_n)$, the state transition equation is shown as follows:

$$M_{i,j} = \big(M_{i,j-1} + M_{i-1,j} + M_{i-1,j-1}\big) k\big(X_i, Y_j\big)$$
$$k\big(X_i, Y_j\big) = e^{-\|X_i - Y_j\|^2}. \tag{8}$$

And the initial state is

$$\begin{cases} M_{i,0} = 0, & i = 1, 2, \ldots, m \\ M_{0,j} = 0, & j = 1, 2, \ldots, n \\ M_{0,0} = 1. \end{cases} \tag{9}$$

Based on the state transition equation and initial state, the global alignment kernel function between X and Y is calculated by

$$K_{GA}(X,Y) = M_{m,n}. \tag{10}$$

With the definitions of MMD and global alignment kernel function, we can define the representativeness of an unlabeled sample as follows.

Definition 6 (Representativeness): Given a PL MTS dataset D and an unlabeled sample $U \in D$, the representativeness of U is calculated by

$$REP(U) = \frac{1}{|\mathcal{L}'|^2} \sum_{X,Y \in \mathcal{L}', X \ne Y} K_{GA}(X,Y) + \frac{1}{|D|^2} \sum_{X,Y \in D, X \ne Y} K_{GA}(X,Y) - \frac{2}{|\mathcal{L}'||D|} \sum_{X \in \mathcal{L}'} \sum_{Y \in D} K_{GA}(X,Y) \tag{11}$$

where $\mathcal{L}' = \mathcal{L} \cup \{U\}$, and $K_{GA}(X,Y) = M_{m,n}$.

C. Sampling Strategy

Sample selection strategy aims to choose the most important samples from unlabeled data for annotation. The ideal sample should have the property that its informativeness should be as large as possible while its representativeness should be as small as possible. For this purpose, we propose an effective bi-objective-based sample selection model which is defined as follows.

Authorized licensed use limited to: Wuhan University. Downloaded on March 18,2021 at 02:14:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Definition 7 (Bi-Objective-Based Sample Selection Model): Given a PL dataset D. To select the most important unlabeled sample for labeling, based on the informativeness and representativeness of unlabeled samples, a bi-objective-based sample selection model is defined as follows:

$$U^* = \text{Arg Min}_{U\in\mathcal{U}} F(U) = \left(\frac{1}{INFO(U)}, REP(U)\right) \quad (12)$$

where INFO(U) and REP(U) represent the informativeness and representativeness of the unlabeled sample U, respectively.

This is a bi-objective optimization problem, and the optimization process of this objective equation is also the selection process of the optimal sample. Here, we introduce two approaches to select the most important unlabeled MTS samples which are shown as follows.

1) Linear Weighted Sum Selection Method: One popular method for bi-objective optimization is to optimize the linear weighted sum of objective functions. So the problem defined in formula (12) could be transformed into a single-objective optimization problem

$$\text{Min}_{\text{s.t. } X\in\mathcal{U}}(1-\alpha)\cdot\frac{1}{INFO(X)} + \alpha\cdot REP(X),\ \ 0\le\alpha\le 1 \quad (13)$$

where parameter α denotes the ratio of importance between INFO(X) and REP(X).

To weaken its sensitivity to α in the process of evaluating unlabeled samples, INFO(X) and REP(X) are normalized in advance. Based on formula (13), all unlabeled samples are ordered and the optimal one is selected for labeling.

2) Indicator-Based Selection Method: The other selection method is to rank informativeness and representativeness in parallel. Motivated by [45], we introduce an indicator-based selection method to obtain the most important unlabeled samples. Since we have two sorting metrics, namely, informativeness and representativeness, for an unlabeled sample X in U, we design two key indicators, $I_1(X)$ and $I_2(X)$, for different comparison preferences

$$I_1(X) = \sum_{Y\in\mathcal{U}, Y\neq X} -e^{-I_e(X,Y)/0.05} \quad (14)$$

$$I_e(X,Y) = \min\left\{\frac{1}{INFO(X)} - \frac{1}{INFO(Y)}, REP(X) - REP(Y)\right\}$$

$$I_2(X) = \min_{X,Y\in\mathcal{U},\ Y \text{ precedes } X}\{I_{SD}(X,Y)\} \quad (15)$$

where Y precedes X means Y is ahead of X in the rank, and

$$I_{SD}(X,Y) = \sqrt{sd\left(\frac{1}{INFO(X)}, \frac{1}{INFO(Y)}\right)^2 + sd(REP(X), REP(Y))^2}$$

$$sd(a,b) = \begin{cases} b-a, & \text{if } a<b \\ 0, & \text{otherwise.}\end{cases}$$

The indicator-based sorting method is shown in Algorithm 1. First, a value γ is randomly generated in the interval of 0 to 1 (line 3). Next, the value γ is compared with the parameter β (line 4). If $\gamma < \beta$, unlabeled samples are ranked in ascending order according to $I_1(\cdot)$ (lines 5, 6, and 7). Otherwise, they are ranked in ascending order according to $I_2(\cdot)$ (lines 9, 10, and 11). Based on different values of the parameter β, we can control the proportions of two indicators in the algorithm. If β is greater than 0.5, the algorithm prefers to exchange a sample with bigger $I_1(\cdot)$ to the latter position, then the query step mainly depends on indicator $I_1(\cdot)$. Otherwise, the indicator $I_2(\cdot)$ is more important when β is smaller than 0.5. Finally, the algorithm returns the best sample in the unlabeled dataset $\mathcal{U}$.

Algorithm 1 Indicators-Based Sorting Algorithm

Require:
Input: A unlabeled dataset $\mathcal{U} = \{X_0, X_1, \ldots X_k, \ldots, X_{|\mathcal{U}|}\}$
Indicators $I_1(\cdot)$ and $I_2(\cdot)$
Parameter β
Ensure
1. **For** $i \leftarrow 0$ to $|\mathcal{U}|/2$ **do**
2. **For** $j \leftarrow |\mathcal{U}| - 1$ to 0 {
3. Take a random value $\gamma \in (0, 1)$
4. **If** $\gamma < \beta$ **then** {
5. **If** $I_1(X_j) > I_1(X_{j+1})$ **then**
6. $swap(X_j, X_{j+1})$ }
7. **Else** {
8. **If** $I_2(X_j) > I_2(X_{j+1})$ **then**
9. $swap(X_j, X_{j+1})$ }}
10. **Output:** the best sample X_b in sorted $\mathcal{U}$

IV. Cost-Effective Crowd Labeling Approach

As we mentioned above that oracles generally have different accuracy and cost due to their uneven expertise. The possibility exists that a labeler with a lower cost can provide a more accurate labeling on some specific samples. Moreover, the decision making by swarm intelligence of multiple oracles is competitive compared with any one of these oracles. Enlightened by this idea, in this section we advance a cost-effective crowd labeling (CCL) approach.

A. Cost-Effective Crowd Labeling Model

The cost of one labeler is usually constant whether the unlabeled sample is in his familiar field or not, so we can choose multiple oracles with the minimum cost while guarantee that the accuracy of labeling meets users' need. To illustrate this idea, we give a simple sample.

Example 1: Given a dataset which can be partitioned into three groups, A, B, and C, as shown in Fig. 1. There are three labelers L1, L2, and L3 with different accuracies on group A, B, and C. Suppose L1 can label samples in A group with a high accuracy while not being good at labeling samples in groups B and C. L2 is good at labeling groups A and B. L3 works well on groups B and C but fails on group A. Meanwhile, the costs of different labelers are also different, and we assume cost (L1) <cost (L2) <cost (L3). For an unlabeled sample U in group A, we notice the best labeler is L1 since labeler L1 could provide high accuracy for labeling U with minimum cost.

Before introducing CCL model, we first give the concepts of labeling confidence and cost of swarm intelligence as following.

Authorized licensed use limited to: Wuhan University. Downloaded on March 18,2021 at 02:14:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

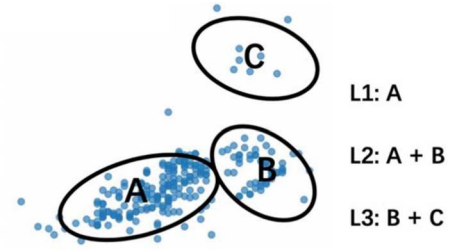

Fig. 1. Example of cost-effective labeling an unlabeled instance.

Definition 8 (Labeling Confidence of Swarm Intelligence): The labeling confidence means the probability of the label given by the crowd being consistent with the ground-truth label of the instance. Given an unlabeled sample U in $\mathcal{U}$ and multiple oracles $\mathcal{H}$ which consistently label the sample U as the same class. Based on decision making of swarm intelligence with these labelers, the labeling confidence of swarm intelligence for the unlabeled sample U is calculated by

$$\text{Confid}(U,\mathcal{H}) = \sum_{i=1}^{m} q_i - \sum_{1\le i<j\le m} q_i q_j + \sum_{1\le i<j<k\le m} q_i q_j q_k + \cdots + (-1)^{m-1} q_i q_j \cdots q_m \tag{16}$$

where the scale of $\mathcal{H}$ is $m = |\mathcal{H}|$, and q_i is the labeling accuracy of the sample U by the ith labeler in $\mathcal{H}$.

Definition 9 (Labeling Cost of Swarm Intelligence): Labeling cost is the cost of each labeler for labeling. Here, we consider different people with varying level of expertise have different costs. Given an unlabeled sample U in $\mathcal{U}$, its class is labeled by swarm intelligence of multiple oracles $\mathcal{H}$ with the scale of $m = |\mathcal{H}|$. The labeling cost of swarm intelligence for the unlabeled sample U are calculated by

$$\text{Cost}(U,\mathcal{H}) = \sum_{i=1}^{m} \text{cost}(a_i) \tag{17}$$

where $\text{cost}(a_i)$ is the cost of oracle a_i in $\mathcal{H}$. Our task is to find an optimal subset of oracles for swarm intelligence labeling with minimum cost and satisfactory accuracy. This optimization problem can be formalized as a cost-effective crowd labeling model.

Definition 10 (Cost-Effective Crowd Labeling): Given the crowd $\mathcal{H}$, an unlabeled MTS sample U, and oracles with different accuracy and costs. Then, the issue of discovering an optimal subset of oracles for swarm intelligence labeling is formulated as

$$\mathcal{H}^* = \text{Arg}\ \min_{\mathcal{H}^*\subseteq\mathcal{H}} \sum_{a\in\mathcal{H}^*} \text{Cost}(a) \quad \text{s.t.} \quad \text{Confid}\left(U,\mathcal{H}^*\right) \ge \delta \tag{18}$$

where $\text{Cost}(a)$ represents the labeling cost of the oracle a, and $\text{Confid}(U,\mathcal{H}^*)$ is defined in formula (16).

B. Crowd Selection Solution

Based on the constrained optimization model in (18), we advance an ALS, and its overview is provided in Algorithm 2. The input of ALS is an unlabeled MTS sample U, accuracy, and costs of all labelers in the crowds. ALS first generates an initialized population of some possible solutions. Then, it uses crossover and mutation operations to generate new possible solutions. Next, all individuals are evaluated based on the evaluation criterion, and the best one of the population is added to the elite pool in each generation. The crossover and mutation operations will be repeated until the stopping conditions are satisfied. The output of ALS is the optimal subset of oracles for swarm intelligence labeling with minimum cost and satisfied confidence.

Algorithm 2 ALS

Require:
Input: An unlabeled sample U
Accuracies and costs of all labelers in the crowds
Confidence threshold δ
Ensure:
1. Randomly generate a population Θ with N individuals using Eq. (19);
2. **Repeat**
3. Evaluate the fitness $F(\mathcal{H})$ for each individual $\mathcal{H}$ using Eq. (18) with respect to the population Θ;
4. Apply crossover operator using Eq. (22);
5. Apply the mutation operator using Eq. (23);
6. Until the stopping criterion is satisfied;
Output: the most suitable subset of all available labelers and its confidence.

The detailed description of the main steps is given as follows.

Initialization: A population Θ with N individuals is initialized randomly (line 1). Each individual in the population Θ consists of a set of binary codes

$$\mathcal{H} = <w^1, w^2, \ldots, w^i, \ldots, w^n>, w^i \in \{0,1\} \tag{19}$$

where n is the number of all available labelers, and each value w^i denotes the i_{th} labeler is selected when $w^i = 1$, otherwise $w^i = 0$.

Evaluation Criterion: The fitness (labeling confidence and cost) of all individuals in the population Θ is evaluated in this step (line 3). For an individual H in the population Θ, its fitness is calculated by

$$F(\mathcal{H}) = \begin{cases} -\text{Cost}(\mathcal{H},U), & \text{if } g(U,\mathcal{H}) \le 0 \\ -\text{Cost}_{\max} - g(U,\mathcal{H}), & \text{if } g(U,\mathcal{H}) > 0 \end{cases} \tag{20}$$

where $\text{Cost}_{\max}$ is the maximum cost among all feasible solutions, and penalty function $g(U,\mathcal{H})$ is calculated as

$$g(U,\mathcal{H}) = \delta - \text{Confid}(U,\mathcal{H})$$

where δ is the confidence threshold, $\text{Confid}(U,\mathcal{H})$ is the labeling confidence of U with the crowds $\mathcal{H}$.

Crossover: Given two selected individuals $\mathcal{H}_a$ and $\mathcal{H}_b$ represented as

$$\mathcal{H}_a = <{w_a}^1, {w_a}^2, {w_a}^3, {w_a}^4, \ldots, {w_a}^n> \quad \mathcal{H}_b = <{w_b}^1, {w_b}^2, {w_b}^3, {w_b}^4, \ldots, {w_b}^m>. \tag{21}$$

Suppose the crossover point of $\mathcal{H}_a$ is between ${w_a}^3$ and ${w_a}^4$ while the crossover point of $\mathcal{H}_b$ is between ${w_b}^3$ and ${w_b}^4$.

Authorized licensed use limited to: Wuhan University. Downloaded on March 18,2021 at 02:14:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Algorithm 3 Framework of Active Semi-Supervised Learning

Require:
Input: Dataset D with two subsets $\mathcal{L}$ and $\mathcal{U}$
Ensure:
1. **Do**
2. Select the most important unlabeled sample U from $\mathcal{U}$ using proposed sampling strategies in Section III;
3. Select the multiple labelers using ALS algorithm;
4. Confidently label the unlabeled sample $\boldsymbol{U}$ using Eq. (18) and add U into $\mathcal{L}$;
5. Automatically classify the *R1NN* sample U^* of U with semi-supervised learning and add U^* into $\mathcal{L}$;
6. **While** (stopping criterion is not satisfied);
7. **Output:** the labeled dataset $\mathcal{L}$.

After the crossover of two individuals $\mathcal{H}_a$ and $\mathcal{H}_b$, the new individuals $\mathcal{H}'_a$ and $\mathcal{H}'_b$ are presented as

$$\begin{aligned}\mathcal{H}'_a &= < {w_b}^1, {w_b}^2, {w_b}^3, {w_a}^4, \ldots, {w_a}^n > \\ \mathcal{H}'_b &= < {w_a}^1, {w_a}^2, {w_a}^3, {w_b}^4, \ldots, {w_b}^m >. \end{aligned} \tag{22}$$

ALS compares the values of the fitness $F(\mathcal{H}_a), F(\mathcal{H}_b), F(\mathcal{H}'_a)$, and $F(\mathcal{H}'_b)$. The better two among four individuals $\mathcal{H}_a$, $\mathcal{H}_b$, $\mathcal{H}'_a$, and $\mathcal{H}'_b$ will be preserved, and the others will be eliminated.

Mutation: The mutation operator randomly selects a locus of an individual $\mathcal{H}_a$ and changes its value of this locus. For instance, the individual $\mathcal{H}_a$ is presented as

$$\mathcal{H}_a = < {w_a}^1, {w_a}^2, {w_a}^3, {w_a}^4, \ldots, {w_a}^n >.$$

Suppose ${w_a}^2$ of individual $\mathcal{H}_a$ at the locus 2 is to be mutated. A new individual $\mathcal{H}'_a$ will be generated as

$$\mathcal{H}'_a = < {w_a}^1, \overline{{w_a}^2}, {w_a}^3, {w_a}^4, \ldots, {w_a}^k > \tag{23}$$

$\overline{{w_a}^2}$ represents the opposite of ${w_a}^2$. That is to say, $\overline{{w_a}^2} = 1$ if ${w_a}^2 = 0$ and $\overline{{w_a}^2} = 0$ if ${w_a}^2 = 1$.

The better one among $\mathcal{H}_a$ and $\mathcal{H}'_a$ will be kept in ALS, and the other will be discarded.

Stopping Criterion: When the maximum number of iterations is reached or the best one has not changed during several iterations, the output is given as the best individual $\mathcal{H}^*$ with the highest fitness.

V. Framework of Active Semi-Supervised Learning

A. Framework

Based on the R1NN method, our active semi-supervised learning framework is presented in algorithm 3. First, the most valuable unlabeled sample U from $\mathcal{U}$ is chosen according to the sample selection model in line 2. Then, the optimal multiple labelers are selected from the crowd using the ALS algorithm in line 3. Next, the label of U is obtained using (18) and U is put into $\mathcal{L}$ in line 4. To enlarge the labeled data, the R1NN sample U^* of U is automatically classified to the same class as U with semi-supervised learning in line 5. The process of active semi-supervised learning continues until the stopping criterion is satisfied. Eventually, the updated labeled dataset $\mathcal{L}$ is obtained in line 7.

B. Stopping Criterion

Our goal is to obtain a high-quality labeled training dataset with minimum cost, so the active learning process should stop when newly labeled samples no longer improve the performance of classification. Therefore, an appropriate stopping criterion should balance the resource cost and the performance of the classifier.

During the specified continuous iterations in active learning, when the variation of the chosen samples' utilities fluctuates in a small scope, it means in the remaining unlabeled data, there's no more important samples to further improve the performance of classification, so we could stop the active learning process. The stopping criterion based on this idea is defined as follows:

$$\begin{aligned} &\left|\max_{i\le k\le j}\text{Utility}(S_k) - \min_{i\le m\le j}\text{Utility}(S_m)\right| \le \varepsilon \\ &\text{s.t. } |i-j| \le \varphi \end{aligned} \tag{24}$$

where S_k is the kth chosen unlabeled sample in the process of active learning; Utility(S_k) represents the score of S_k, which is calculated by linear weighted sum selection (LWS) or indicator-based selection (IBS) method; $|\mathcal{U}|$ is the size of $\mathcal{U}$; the parameter ε is a stable amplitude, and here it is set to 0.001. φ is a stable interval, which is set to $0.1|\mathcal{U}|$.

C. Evaluation Criterion

To illustrate the efficiency of our proposed method, we adopt the 1-NN method with DTW similarity to evaluate the classification performance based on the labeled training data, which is obtained with the proposed active semi-supervised learning.

In the validation step, we employ two widely used criteria, F-measure and accuracy, to measure the performance of classification results. The F-measure is defined as $F = 2 * p * r/(p + r)$, where p and r represent the precision and recall of classification, respectively. F-measure is larger when both of its precision and recall are higher. High F-measure and accuracy value means good classification performance.

D. Complexity Analysis

The whole algorithm can be divided into two stages, including the sampling stage and the labeling stage.

In the sampling stage, suppose the size of MTS data is M, and an MTS contains m variables. For the informativeness, whose calculation is based on RkNN and kNN, the complexity is $O(M^2)$. Representativeness is based on MMD and its complexity is decided by a kernel matrix whose complexity is $O(m^2M^2)$. The complexity for IBS and LWS is O(M). So the total complexity for the sampling stage is $O(m^2M^2)$.

For the labeling stage, we choose N labelers for annotation from the crowds with the size n, then the complexity for ALS is $O(Nn^2)$.

As a result, for the whole algorithm, suppose we choose t unlabeled samples from the whole dataset, then the algorithm need to iterate t times. The calculation of RkNN, kNN, and the kernel matrix are only done once at the beginning of the algorithm. So the total complexity of the algorithm is $O(m^2M^2 + tNn^2)$.

Authorized licensed use limited to: Wuhan University. Downloaded on March 18,2021 at 02:14:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
DESCRIPTIONS OF DATASETS USED IN EXPERIMENTS

	Classes	Variables	Max length	Min length	Size
JV	2	12	29	7	581
WG	2	3	315	315	1120
ECG	2	2	152	39	199
WAFER	2	6	198	104	274
PB	2	2	8	8	1200
CT	2	3	205	100	1200
REF	2	6	15	15	463
BCI	2	28	500	500	179
UGL	2	3	310	310	201
ASL	2	22	90	47	639

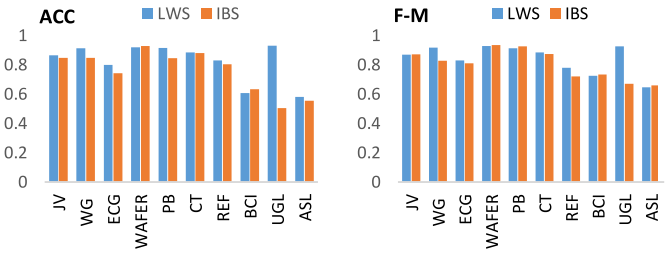

Fig. 2. Experiment results of 1-NN classification method on 10 datasets when 5 percent of the original unlabeled data is labeled and added to the labeled training data with two proposed sampling methods.

VI. EXPERIMENTS

To identify the performance of our proposed methods, we evaluate our method and comparison methods with 10 real-world datasets. For the training data of each dataset, in the beginning, there is only one positive sample in P and the rest of the samples are treated as U. We randomly sample the labeled sample for initialization and repeat the training process ten times to mitigate the initiation sensitivity. Finally, the F-measure and accuracy of the classifier are averaged over all runs.

The 10 real-world datasets are Japanese Vowels (JVs), uWaveGestureLibrary (WG), ECG, Wafer, Pen Digits (PB), character trajectories (CTs), robot execution failures (REFs), Berlin brain–computer interface (BCI), UGL and Australian sign language sighs (ASL) datasets [46]–[48]. The description of these datasets are shown in Table II. Here, we just deal with the issue of two-class data. If the data is more than two classes, we group the former half of these classes as positive and the other half as negative.

A. Experiment Setting

To simulate the labelers, for all datasets we adopt a density peak-based clustering method to divide each one into several domains [49]. The accuracy of a labeler on a domain is set in the range of [0.5, 0.6] or [0.6, 0.7] randomly. Each domain of all datasets is randomly assigned an accuracy range for each labeler. Then, based on the normal distribution and its 3σ principle, in each domain of all datasets a certain accuracy is assigned to a labeler. Based on the mean accuracy of a labeler a_i on a dataset D, the cost of this labeler is set as follows:

$$\text{Cost}(a_i) = \gamma^{\eta \cdot \text{Mean\_Acc}(a_i, D)}$$

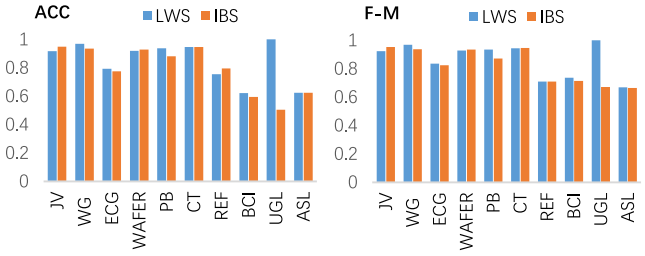

Fig. 3. Experiment results of 1-NN classification method on 10 datasets when 10 percent of the original unlabeled data is labeled and added to the labeled training data with two proposed sampling methods.

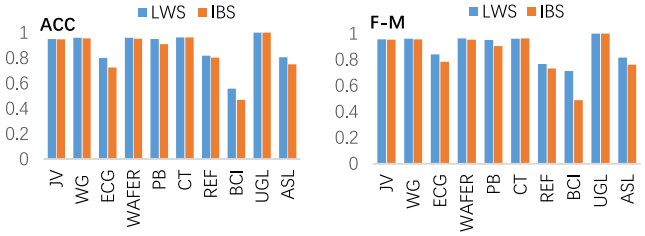

Fig. 4. Experiment results of 1-NN classification method on 10 datasets when 20 percent of the original unlabeled data is labeled and added to the labeled training data with two proposed sampling methods.

where Mean_Acc(a_i, D) is the mean accuracy of a_i on dataset D, $\gamma = 3$ and $\eta = 20$. The accuracy and costs stay unchanged during all experiments after setting.

In the two proposed sampling methods, the parameter α in the LWS method is used to adjust the importance of informativeness and representativeness in the evaluation of an unlabeled sample. It was found that $\alpha \in [0.4, 0.8]$ works well because INFO(X) and REP(X) in formula (13) are normalized and LWS is not sensitive to the parameter α. For the simplicity, in experiments, the parameter β is set to 0.3 for all datasets. The parameter β in the IBS method is used to adjust the proportion of two indicators and its optimal value is data-dependent. To make a fair comparison with the LWS method, the parameter β is set to the optimal value for a specific dataset.

B. Analysis of Proposed Sampling Approach

1) Effectiveness of Two Proposed Sampling Methods: To compare the effectiveness of our proposed two sampling strategies in Section II, here we evaluate the classification performance of 1-NN classifier with an absolutely accurate labeler, and one data sampled in each iteration. For brevity, the proposed sampling methods are abbreviated as follows.

1) *IR-LWS:* Informativeness and representativeness sampling with Linear Weighted Sum method.
2) *IR-IBS:* Informativeness and representativeness-based sampling with Indicator-Based Selection method.

To better show the classification performance as the scale of the labeled training data gradually grows, different percentages of unlabeled samples are labeled and added to the labeled training data in the experiments. Here, we just show the classification results when 5, 10, and 20 percent of the original unlabeled data are labeled and added to the labeled

Authorized licensed use limited to: Wuhan University. Downloaded on March 18,2021 at 02:14:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

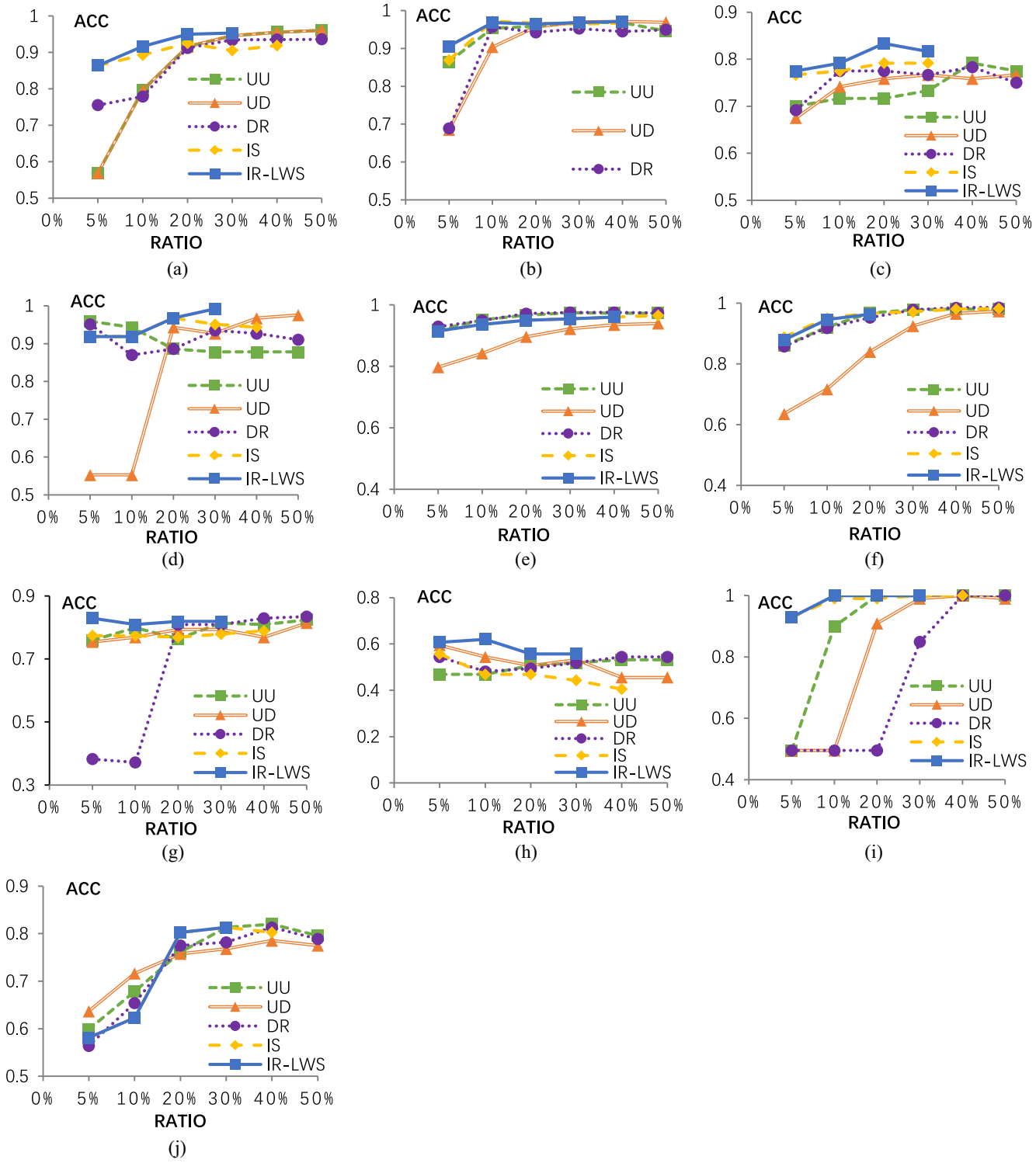

Fig. 5. Experiment results of 1-NN classification method on 10 datasets, where different ratios of unlabeled samples are chosen with five sampling methods and added to the labeled training data. (a) JV. (b) WG. (c) ECG. (d) WAFER. (e) PB. (f) CT. (g) REF. (h) BCI. (i) UGL. (j) ASL.

training data, which are plotted in Figs. 2–4, respectively. They show IR-LWS is more competitive than IR-IBS on all datasets. For instance, when 5, 10, and 20 percent of unlabeled samples are annotated, the accuracies of IR-LWS on JV, WG, and ECG datasets are (0.01, 0.0, 0.06), (0.04, 0.03, 0.01), and (0, 0.01, 0.07) higher than those of IR-IBS, respectively. Similarly, the F-measures of IR-LWS on the three datasets with three sampling percentages are higher than those of IR-IBS by (0, 0.09, 0.02), (−0.04, 0.01, 0.01), and (0.01, 0, 0.06), respectively.

2) Comparison With Other Sampling Methods: To further investigate the effectiveness of IR-LWS, here we compare IR-LWS with state-of-the-art methods based on 1-NN classifier with an absolutely accurate labeler, and one data sampled in each iteration. The comparison methods are abbreviated as follows.

Authorized licensed use limited to: Wuhan University. Downloaded on March 18,2021 at 02:14:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

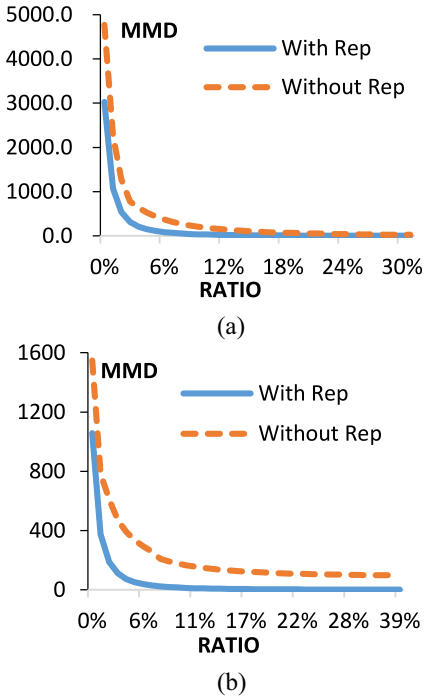

Fig. 6. MMD distribution difference visualization. (a) JV. (b) WG.

1) *IS:* Informativeness-based sampling method without representativeness [7].
2) *UU:* A sampling method based on uncertainty and utility proposed in [7].
3) *UD:* A sampling method based on uncertainty and density in [50].
4) *DR:* A sampling method based on in terms of density-based reranking in [50].

Here, we use these sampling strategies with the proposed semi-supervised learning method to obtain the same number of labeled training data, which is used to learn a classification model. For the labeling strategy, we adopt the popular method, which is just a labeler with absolutely accuracy. To show the changes in the performance of classification with the manually labeled samples being gradually augmented, we do experiments on different percentage of manually labeled samples.

Fig. 5 shows 1-NN classification performance of the proposed active semi-supervised methods with the above sampling methods on all datasets, respectively. In Fig. 5, it shows that when the number of manually labeled samples is smaller, the classification performance of IR-LWS is far better than other methods. Among four comparison methods, the method whose performance is closest to IR-LWS is IS, but the overall performance of IR-LWS is better than that of IS. When the percentage of labeling samples becomes larger, the classification performances of all sampling method gradually become similar. It is reasonable because the number of manually labeled samples is larger, most valuable samples have been labeled and added to the labeled training data by all sampling methods. Apparently, IR-LWS is more competitive for data sampling, which could obtain a high-quality dataset to learn a good classification model when the expert resource is limited.

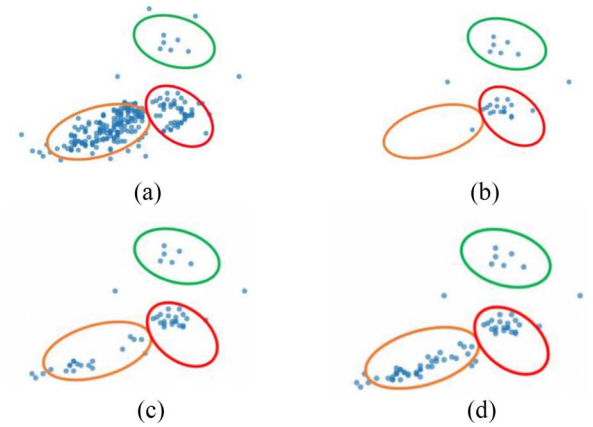

Fig. 7. Effect of IR-LWS sampling method on JV dataset. (a) 100%. (b) 10%. (c) 20%. (d) 30%.

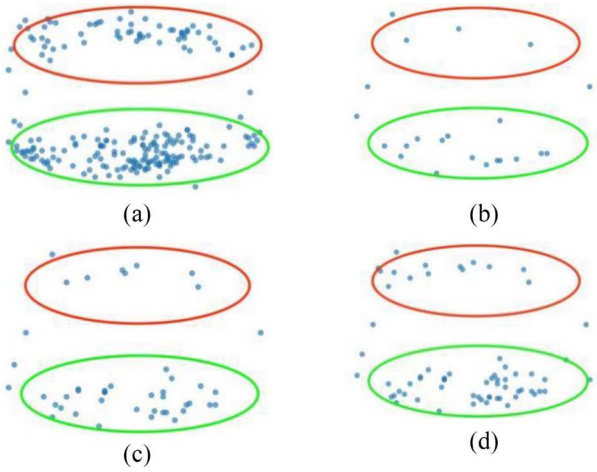

Fig. 8. Effect of IR-LWS sampling method on WG dataset. (a) 100%. (b) 10%. (c) 20%. (d) 30%.

At the same time, from Fig. 5 we can see the trends of accuracy and F-measure of IR-LWS consistently increase on most datasets as more unlabeled samples are annotated and added to the training data which means the classification performance of IR-LWS is more stable than that of other sampling methods. The reason behind it is that IR-LWS can select better unlabeled samples that are more helpful for the classification.

Moreover, we analyze the number of unlabeled samples being annotated when the active semi-supervised learning process ends. From Fig. 5 we see that the process of IR-LWS ends with a smaller number of annotated samples. This is because IR-LWS could select more valuable unlabeled samples, and the chosen ones could help other informative samples be annotated quickly in the process of semi-supervised learning, resulting in an earlier ending than other methods.

3) Effect of Representativeness in IR-LWS: To investigate the effect of representativeness in the IR-LWS sampling method, in this section we compare it with informativeness-based sampling. To facilitate comparison, they are abbreviated as follows.

1) *With Rep:* our proposed IR-LWS.
2) *Without Rep:* Informativeness-based sampling without considering representativeness.

Authorized licensed use limited to: Wuhan University. Downloaded on March 18,2021 at 02:14:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

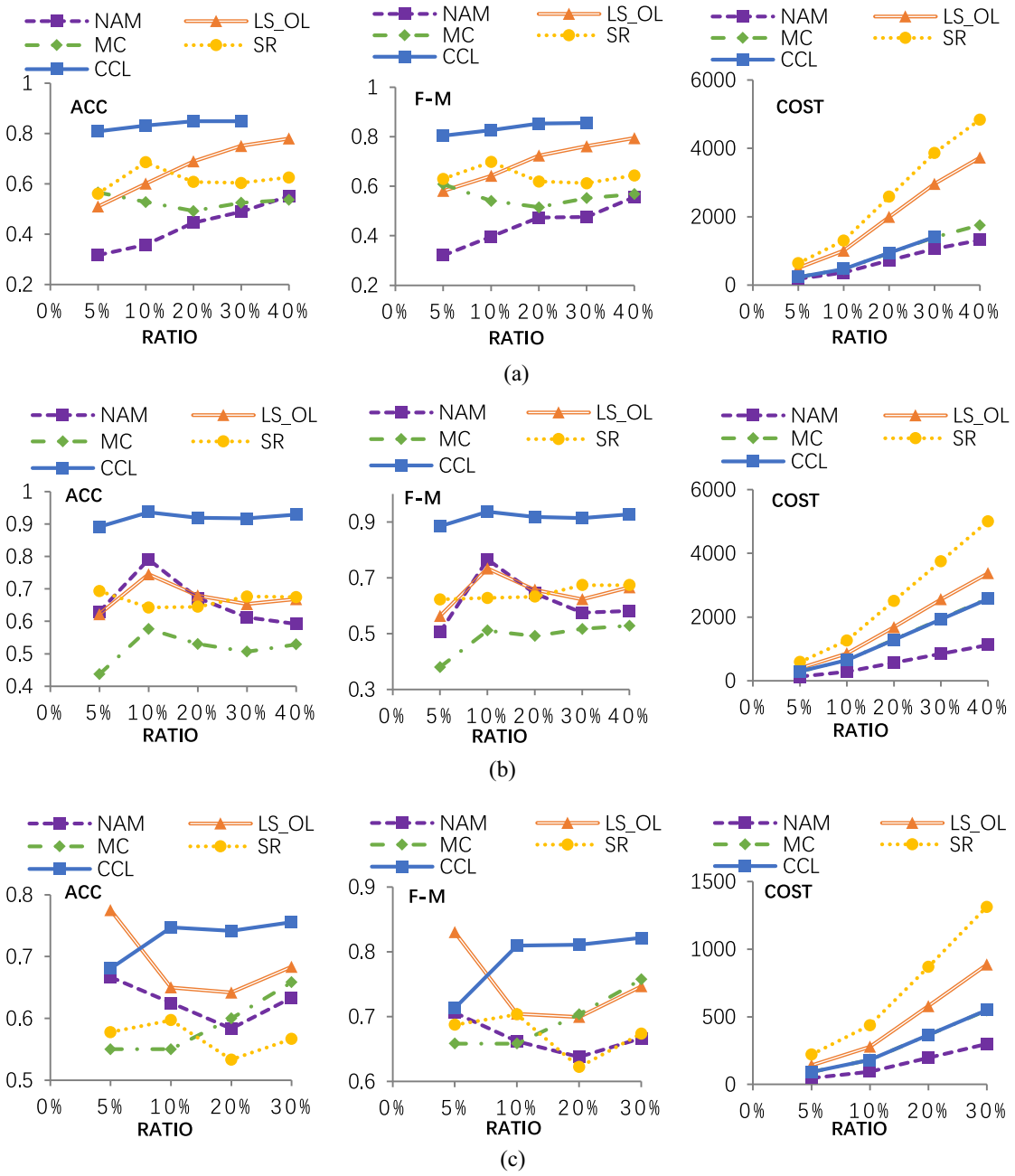

Fig. 9. Experiment results of 1-NN classification method on JV, WG, and ECG datasets, where different ratios of unlabeled samples are labeled and added to the training data by 5 labeling methods. (a) JV. (b) WG. (c) ECG.

As we have introduced in the calculation of representativeness, MMD can evaluate the distribution discrepancy of two datasets. The smaller the distribution discrepancy, the more similar the spatial distribution of two datasets. So if the MMD is lower, that means the labeled dataset obtained by the method has a more similar distribution with the whole dataset and can better represent the whole dataset. As a result, to show the effectiveness of representativeness, the MMD values between labeled training data and the whole data (including labeled and unlabeled data) are compared as unlabeled samples being continually added to the labeled data. Here, only the experiment results of JV and WG datasets of are shown in Fig. 6, and IR-LWS works on other datasets in a similar way. They show that the IR-LWS sampling method could effectively enhance the diversity of the labeled training data by selecting some classical unlabeled sample to the labeled data, which consider the spatial distribution of the whole data.

For instance, when the number of chosen unlabeled samples reaches 6% of the whole data on JV and WG datasets, comparing two MMDs we see that the MMDs of IR-LWS are 287 and 318 less than those of informativeness-based sampling method, respectively. A smaller MMD means a more similar spatial distribution between the two datasets. Therefore, representativeness effectively improves the diversity of the labeled training data in the process of instance selection of active learning.

Authorized licensed use limited to: Wuhan University. Downloaded on March 18,2021 at 02:14:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE III
ACCURACY, F-MEASURE, AND COST OF VARIOUS LABELING METHODS ON 10 DATASETS, WHERE 10 PERCENT OF UNLABELED EXAMPLES ARE LABELED AND ADDED TO THE TRAINING DATA

	ACC					F-M					COST				
	NAM	LS_OL	MC	SR	CCL	NAM	LS_OL	MC	SR	CCL	NAM	LS_OL	MC	SR	CCL
JV	0.36	0.60	0.53	0.69	0.83	0.40	0.64	0.54	0.70	0.83	364.14	1007.38	467.56	1303.95	469.55
WG	0.79	0.75	0.58	0.64	0.94	0.77	0.73	0.51	0.63	0.94	286.60	860.55	645.00	1260.86	643.78
ECG	0.63	0.65	0.55	0.60	0.75	0.66	0.70	0.66	0.70	0.81	94.80	278.69	182.22	437.99	180.85
WAFER	0.49	0.49	0.72	0.68	0.89	0.26	0.26	0.73	0.71	0.90	224.82	378.09	147.45	390.23	178.07
PB	0.69	0.81	0.73	0.71	0.92	0.68	0.80	0.72	0.70	0.91	872.29	6499.08	4800.48	2551.16	1717.18
CT	0.69	0.72	0.62	0.70	0.92	0.69	0.71	0.63	0.70	0.92	903.51	3078.62	2009.11	4124.07	1968.44
REF	0.56	0.71	0.59	0.66	0.75	0.32	0.55	0.59	0.50	0.62	366.98	976.77	561.11	1853.51	723.77
BCI	0.44	0.53	0.49	0.49	0.49	0.42	0.49	0.57	0.53	0.57	108.14	358.94	185.51	516.45	207.52
UGL	0.82	0.82	0.58	0.84	0.89	0.82	0.82	0.48	0.84	0.87	122.00	250.44	151.17	449.12	165.31
ASL	0.57	0.47	0.44	0.54	0.59	0.61	0.53	0.42	0.52	0.61	560.39	1857.32	954.86	2770.90	697.62

Furthermore, we visualize the effect of representativeness for the IR-LWS method. Since the dimension of MTS data is very high, we adopt the DTW distance between two samples to transform the whole MTS data into 2-D space. Figs. 7(a) and 8(a) show the spatial distributions of the whole JV and WG datasets. Fig. 7(b)–(d) show the spatial distributions of labeled JV data when the ratio of labeled data increases to 10%, 20%, and 30%, respectively. Fig. 7(b)–(d) show that as the chosen samples are increasingly added to the labeled data, the spatial distributions of the labeled data are more and more similar to that of the whole data. Similarly, Fig. 8 shows the same phenomenon. The reason is that the proposed sample selection method effectively considers the spatial distribution of data, which could enhance the diversity of the labeled data.

C. Analysis of Proposed Labeling Strategy

To investigate the performance and cost of our proposed labeling strategy, we conduct experiments to compare with 2 state-of-the-art labeling methods, NAM [19] and LS_OL [32], and two methods designed by us, MC, and SR. The detail of the four comparison labeling methods are as follows.

1) *NAM:* Selecting the labeler with the best accuracy for a dataset.
2) *LS_OL:* Selecting k labelers with the best accuracy for a dataset. In the process of labeling an sample, the labeler with the best accuracy among these k labelers is further chosen to determine its class according to its specific domain.
3) *MC:* Selecting k labelers with the lowest cost for a dataset. In the process of labeling an sample, the labeler with the best accuracy among these k labelers is further chosen to determine its class according to its specific domain.
4) *SR:* Selecting k labelers with the lowest cost. In the process of labeling an sample, the labeler with best accuracy among these k labelers is further chosen to determine its class according to its specific domain.
5) *CCL:* Our proposed CCL approach.

To be fair, in experiments, the proposed IR-LWS sampling method is used for instance selection. And the parameter k in other labeling methods is set to 5 since CCL selects multiple labelers to improve the accuracy of labeling. To verify the performance of CCL, the analysis process is divided in two steps. First, we perform a classification comparison using the 1-NN classification method, and the labeled training data is gradually augmented as more and more unlabeled samples are labeled by different strategies. Next, we compare the labeling cost when different percentages of unlabeled samples are annotated. We conduct experiments on 10 datasets and the results are all similar. Here, only the results of JV, WG, and ECG datasets are presented in Fig. 9. The accuracy, F-measure, and COST of various labeling methods on 10 datasets are listed in Table III.

Fig. 9 shows that when the number of labeled samples becomes larger, the classification performance of CCL is steadily getting better, and outperforms other labeling methods, which means CCL can confidently label the unlabeled samples and effectively improve the quality of labeled training data as well as the classification performance. For instance, when the number of labeled samples reaches 10% of the whole data on JV, WG, and ECG datasets, the accuracies of CCL are larger than those of SR, NAM, LS_OL, and MC by (0.14, 0.30, 0.14), (0.04,0.31, 0.25), (0.23, 0.19, 0.09), and (0.30, 0.36, 0.19), respectively, and the F-measures of CCL are (0.13,0.31, 0.11), (0.43, 0.17, 0.15), (0.19, 0.23, 0.11), and (0.29, 0.43, 0.15) higher than those of SR, NAM, LS_OL, and MC, respectively. This means a higher quality data is obtained by CCL and the core reason is that CCL adopts swarm intelligence to label unlabeled samples with crowds, which could improve the quality of labeled training data, and further learn a good classification module. In comparison, other labeling methods just use one labeler for labeling, so the labeled training data generated by these labeling methods is worse in quality. Meanwhile, the costs of five labeling strategies on three datasets are also compared in Table III. We can see that generally speaking, selecting multiple labelers results in higher accuracy and F-measure, because it can take the advantage of crowd intelligence, but cost is higher, too. it shows that the proposed method has the best performance of accuracy and F-measure on most datasets. And the cost is only a bit higher than NAM which only chooses one labeler. Specifically, when the ratio of labeled samples reaches 10% on JV, WG,

Authorized licensed use limited to: Wuhan University. Downloaded on March 18,2021 at 02:14:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

and ECG datasets, the costs of CCL are 469, 643, and 180, respectively. Compared with SR, LS_OL, and MC, the costs of CCL are lower than those of these methods by (834, 617, 257), (538, 217, 98), and $(-2, -2, 2)$, respectively. The reason is that CCL could select those labelers with the lowest cost to confidently label samples. This means our method is able to find labelers with lower cost while preserving the labeling confidence. So our method is very competitive in both classification performance and cost control.

To sum up, experimental results show the remarkable fact that CCL is a cost-effective labeling method, which could select multiple oracles with minimal cost to confidently label unlabeled samples.

VII. Conclusion

To deal with the limitations of conventional active learning algorithms, in this article, we propose a novel framework of active semi-supervised learning, which can obtain a high-quality data with the minimum cost. The new framework has the following properties.

1) Two valid measure criteria, informativeness, and representativeness, are introduced to evaluate the importance of an unlabeled sample from different perspectives. And two sampling strategies are proposed to measure the importance of an unlabeled sample based on the combination of both criteria.
2) A cost-sensitive crowd labeling module is set to formulate the issue of confidently labeling an unlabeled sample with the minimum cost by taking advantage of the swarm intelligence of crowds.

Also, an optimization algorithm is advanced to obtain the multiple labelers, who can confidently identify the class label of an unlabeled sample with minimum cost. In the future, we will continue research on imbalanced unlabeled data to make our algorithm more stable and perform better on realistic problems.

References

[1] K. Yan, Z. Ji, H. Lu, J. Huang, W. Shen, and Y. Xue, "Fast and accurate classification of time series data using extended ELM: Application in fault diagnosis of air handling units," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 7, pp. 1349–1356, Jul. 2019.

[2] S. Deng, B. Wang, S. Huang, C. Yue, J. Zhou, and G. Wang, "Self-adaptive framework for efficient stream data classification on storm," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 1, pp. 123–136, Jan. 2020.

[3] J. Shan H. Zhang, W. Liu, and Q. Liu, "Online active learning ensemble framework for drifted data streams," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 30, no. 2, pp. 486–498, Feb. 2019.

[4] G. He, Y. Pan, X. Xia, J. He, R. Peng, and N. N. Xiong, "A fast semi-supervised clustering framework for large-scale time series data," *IEEE Trans. Syst., Man, Cybern., Syst.*, early access, Aug. 20, 2016, doi: 10.1109/TSMC.2019.2931731.

[5] Y. Li, Y. Wang, D. Yu, Y. Ning, P. Hu, and R. Zhao, "ASCENT: Active supervision for semi-supervised learning," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 5, pp. 868–882, May 2020, doi: 10.1109/TKDE.2019.2897307.

[6] E. Lughofer and M. Pratama, "Online active learning in data stream regression using uncertainty sampling based on evolving generalized fuzzy models," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 1, pp. 292–309, Feb. 2018.

[7] G. He, Y. Li, and W. Zhao, "An uncertainty and density based active semi-supervised learning scheme for positive unlabeled multivariate time series classification," *Knowl. Based Syst.*, vol. 124, pp. 80–92, May 2017.

[8] R. Wang, X. Wang, S. Kwong, and C. Xu, "Incorporating diversity and informativeness in multiple-instance active learning," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 6, pp. 1460–1475, Dec. 2017.

[9] H. Wang, Y. Jin, and J. Doherty, "Committee-based active learning for surrogate-assisted particle swarm optimization of expensive problems," *IEEE Trans. Cybern.*, vol. 47, no. 9, pp. 2664–2677, Sep. 2017.

[10] Y. Yang and M. Loog, "A variance maximization criterion for active learning," *Pattern Recognit.*, vol. 78, pp. 358–370, Jun. 2018.

[11] S. Hao, J. Lu, P. Zhao, C. Zhang, S. Hoi, and C. Miao, "Second-order online active learning and its applications," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 7, pp. 1338–1351, Jul. 2018.

[12] W. Fu, M. Wang, S. Hao, and X. Wu, "Scalable active learning by approximated error reduction," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discover. Data Min.*, London, U.K., 2018, pp. 1396–1405.

[13] Y. Yan and S. Huang, "Cost-effective active learning for hierarchical multi-label classification," in *Proc. 27th Int. Joint Conf. Artif. Intell. Main Track*, Stockholm, Sweden, 2018, pp. 2962–2968.

[14] Z. Qiu, D. Miller, and G. Kesidis, "A maximum entropy framework for semisupervised and active learning with unknown and label-scarce classes," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 28, no. 4, pp. 917–933, Apr. 2017.

[15] J. Lipor, B. Wong, D. Scavia, B. Kerkez, and L. Balzano, "Distance-penalized active learning using quantile search," *IEEE Trans. Signal Process.*, vol. 65, no. 20, pp. 5453–5465, Oct. 2017.

[16] Z. Wang and J. Ye, "Querying discriminative and representative samples for batch mode active learning," *ACM Trans. Knowl. Discover. Data*, vol. 9, no. 3, pp. 158–166, Apr. 2015.

[17] S. J. Huang, J. L. Chen, X. Mu, and Z. H. Zhou, "Cost-effective active learning from diverse labelers," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Melbourne, VIC, Australia, 2017, pp. 1879–1885.

[18] J. Zhang, M. Wu, and V. Sheng, "Ensemble learning from crowds," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 8, pp. 1506–1519, Aug. 2019.

[19] S. Li, Y. Jiang, Nitesh V. Chawla, and Z. Zhou, "Multi-label learning from crowds," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 7, pp. 1369–1382, Jul. 2019.

[20] M. Edoardo, T. Long, and R. Nicholas, "On the efficiency of data collection for crowdsourced classification," in *Proc. 27th Int. Joint Conf. Artif. Intell. Main Track*, Stockholm, Sweden, 2018, pp. 1568–1575.

[21] G. He, Y. Duan, R. Peng, X. Jing, T. Qian, and L. Wang, "Early classification on multivariate time series," *Neurocomputing*, vol. 149, pp. 777–787, Feb. 2015.

[22] S. Mohamad, A. Bouchachia, and M. Sayed-Mouchaweh, "A bi-criteria active learning algorithm for dynamic data streams," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 74–86, Jan. 2018.

[23] W. Huang, M. Yin, J. Li, and S. Xie, "Deep clustering via weighted k-subspace network," *IEEE Signal Process. Lett.*, vol. 26, no. 11, pp. 1628–1632, Nov. 2019.

[24] M. Yin, S. Xie, Z. Wu, Y. Zhang, and J. Gao, "Subspace clustering via learning an adaptive low-rank graph," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3716–3728, Aug. 2018.

[25] M. Yin, J. Gao, Z. Lin, Q. Shi, and Y. Guo, "Dual graph regularized latent low-rank representation for subspace clustering," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4918–4933, Dec. 2015.

[26] J. Murphy and M. Maggioni, "Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1829–1945, Mar. 2019.

[27] C. Yu, J. Hansen, C. Yu, and J. Hansen, "Active learning based constrained clustering for speaker diarization," *IEEE/ACM Trans, Audio, Speech, Language Process.*, vol. 25, no. 11, pp. 2188–2198, Nov. 2017.

[28] S. Huang, R. Jin, and Z. Zhou, "Active learning by querying informative and representative examples," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 1936–1949, Oct. 2014.

[29] D. Wu, "Pool-based sequential active learning for regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1348–1359, May 2019.

[30] B. Du *et al.*, "Exploring representativeness and informativeness for active learning," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 14–26, Jan. 2017.

[31] G. Li, J. Wang, Y. Zheng, and M. J. Franklin, "Crowdsourced data management: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 9, pp. 2296–2319, Sep. 2016.

Authorized licensed use limited to: Wuhan University. Downloaded on March 18,2021 at 02:14:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[32] Y. Liu and M. Liu, "An online learning approach to improving the quality of crowd-sourcing," *IEEE/ACM Trans. Netw.*, vol. 25, no. 4, pp. 2166–2179, Aug. 2017.

[33] K. Atarashi, S. Oyama, and M. Kurihara, "Semi-supervised learning from crowds using deep generative models," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, New Orleans, LA, USA, 2018, pp. 1555–1562.

[34] V. Shengs and J. Zhang, "Machine learning with crowdsourcing: A brief summary of the past research and future directions," in *Proc. 33rd AAAI Conf. Artif. Intell. (AAAI)*, Honolulu, HI, USA, 2019, pp. 9837–9843.

[35] J. Zhang, X. Wu, and V. Shengs, "Active learning with imbalanced multiple noisy labeling," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 1095–1107, May 2015.

[36] W. N. Hsu and H. T. Lin, "Active learning by learning," in *Proc. 29nd AAAI Conf. Artif. Intell. (AAAI)*, Austin, TX, USA, 2015, pp. 2659–2665.

[37] C. Persello, A. Boularias, M. Dalponte, T. Gobakken, E. Næsset, and B. Schölkopf, "Cost-sensitive active learning with lookahead: Optimizing field surveys for remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6652–6664, Oct. 2014.

[38] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2591–2600, Dec. 2017.

[39] K. Wang, L. Lin, X. Yan, Z. Chen, D. Zhang, and L. Zhang, "Cost-effective object detection: Active sample mining with switchable selection criteria," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 834–850, Mar. 2019.

[40] Y. Fu, B. Li, X. Zhu, and C. Zhang, "Active learning without knowing individual instance labels: A pairwise label homogeneity query approach," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 4, pp. 808–822, Apr. 2014.

[41] R. Chattopadhyay, Z. Wang, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, "Batch mode active sampling based on marginal probability distribution matching," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discover. Data Min.*, Beijing, China, 2012, pp. 741–749.

[42] W. Hu, J. Gao, B. Li, O. Wu, J. Du, and S. J. Maybank, "Anomaly detection using local kernel density estimation and context-based regression," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 2, pp. 218–233, Feb. 2020, doi: 10.1109/TKDE.2018.2882404.

[43] C. Chang and H. Huang, "Automatic tuning of the RBF kernel parameter for batch-mode active learning algorithms: A scalable framework," *IEEE Trans. Cybern.*, vol. 49, no. 12, pp. 4460–4472, Dec. 2019. doi: 10.1109/TCYB.2018.2869861.

[44] S. Lin and J. Zeng, "Fast learning with polynomial kernels," *IEEE Trans. Cybern.*, vol. 49, no. 10, pp. 3780–3792, Oct. 2019.

[45] B. Li, K. Tang, J. Li, and X. Yao, "Stochastic ranking algorithm for many-objective optimization based on multiple indicators," *IEEE Trans. Evol. Comput.*, vol. 20, no. 6, pp. 924–938, Dec. 2016.

[46] *UCI Machine Learning Repository*. Accessed: Jul. 31, 2019. [Online]. Available: http://archive.ics.uci.edu/ml/datasets.php

[47] (2004). *Robert T. Olszewski*. [Online]. Available: http://www.cs.cmu.edu/∼bobski/

[48] (2019). *UCR Time Series Classification Archive*. [Online]. Available: https://www.cs.ucr.edu/∼eamonn/time_series_data_2018/

[49] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 3, pp. 1492–1496, Jun. 2014.

[50] J. Zhu, H. Wang, B. K. Tsou, and M. Ma, "Active learning with sampling by uncertainty and density for data annotations," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1323–1331, Aug. 2010.

Guoliang He received the Ph.D. degree in computer software and theory from Wuhan University, Wuhan, China, in 2007.

He is currently an Associate Professor with the School of Computer Science, Wuhan University. His current research interests include data mining, machine learning, and intelligent algorithms.

Bing Li (Member, IEEE) received the Ph.D. degree in the computer science from the Huazhong University of Science and Technology, Wuhan, China, in 2002.

His main research areas are service computing, software engineering, artificial intelligence, complex system, and cloud computing.

Han Wang received the B.S. degree in physics from Wuhan University, Wuhan, China, in 2019. She is currently pursuing the graduation degree with the School of Computer Science.

Her research interest includes data mining and artificial intelligence.

Wenjun Jiang received the B.A. degree in marketing from Northeastern University, Qinhuangdao, China, in 2019. She is currently pursuing the graduation degree with the School of Computer Science, Wuhan University, Wuhan, China.

Her research interest includes data mining and machine learning.

Authorized licensed use limited to: Wuhan University. Downloaded on March 18,2021 at 02:14:40 UTC from IEEE Xplore. Restrictions apply.