IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 20, 2020, accepted September 14, 2020, date of publication September 18, 2020, date of current version September 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3025036

An Active Learning Algorithm Based on Shannon Entropy for Constraint-Based Clustering

DUO WEN CHEN AND YING HUA JIN
School of Applied Mathematics, Guangdong University of Technology, Guangzhou 510520, China

Corresponding author: Ying Hua Jin (jyh@mail.ustc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 11401114, and in part by the Guangdong Province Natural Science Fund under Grant S2012040007622.

ABSTRACT Pairwise constraints could enhance clustering performance in constraint-based clustering problems, especially when these pairwise constraints are informative. In this paper, a novel active learning pairwise constraint formulation algorithm would be constructed with aim to formulate informative pairwise constraints efficiently and economically. This algorithm consists of three phases: Selecting, Exploring and Consolidating. In Selecting phase, some type of unsupervised clustering algorithm is used to obtain an informative data set in terms of Shannon entropy. In Exploring phase, some type of farthest-first strategy is used to construct a series of query with aim to construct clustering skeleton set structure and informative pairwise constraints are also collected meanwhile based on the informative data set. If the number of skeleton sets equals the number of clusters, the new algorithm gets into third phase Consolidating; otherwise, it would finish. In Consolidating phase, non-skeleton points included in the informative data set are used to construct a series of query with skeleton set representative points constructed in Exploring phase. And some type of priority principle is proposed to help collect more must-link pairwise constraints. Treat the well-known MPCK-means (metric pairwise constrained K-means) as the underlying constraint-based semi-supervised clustering algorithm and data experiment comparison between this new algorithm and its counterparts would be done. Experiment outcome shows that significant improvement of this new algorithm.

INDEX TERMS Semi-supervised clustering, pairwise constraint, active learning, entropy, skeleton set.

I. INTRODUCTION

In the domain of machine learning, a traditional unsupervised clustering which classifies samples into different categories uses only similarity between samples [1]. In general, a supervised clustering method has better performance (higher clustering accuracy) than the traditional unsupervised clustering since the former makes the most of some type of prior information [2]. As everyone knows, the process of collecting prior information is time-consuming and costly. Hence, researchers are really interested in how to collect prior information efficiently and economically. The type of pairwise constraints [3] is popular and widely studied. It consists of two categories: must-link and cannot-link. Must-link constraints stipulate that two samples involved must simultaneously belong to some cluster and cannot-link constraints stipulate that two samples involved must belong to two different clusters. Since the type of pairwise constraints contains only must-link and cannot-link type information but not specific sample tag, a supervised clustering which is only involved in pairwise constraints is renamed and called semi-supervised clustering. Previous study [4]–[9] has shown that pairwise constraints could enhance clustering accuracy significantly. However, if pairwise constraints are not properly selected, they may even degrade the clustering performance [10]. The high cost and time consumption may be unaffordable in the process of collecting a large number of pairwise constraints, especially when the inspection to pairwise constraints is manually operated. In order to solve this dilemma, a series of active learning algorithms are proposed with aim to collect pairwise constraints efficiently and economically. The key point in these algorithms how to pick the most informative pairwise constraints and avoid non-informative ones.

Active learning is already studied and used in many fields, such as image processing [11]–[14], text processing [15]–[18] and so on [19], [20], while the research of active learning

The associate editor coordinating the review of this manuscript and approving it for publication was Qilian Liang.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

in semi-supervised clustering based on pairwise constraints is relatively limited [21]. In general, existing methods of selecting pairwise constraints can be roughly divided into two categories: initial selection [9], [22] and iterative selection [21]. The first category is that a set of pairwise constraints is given in advance before executing semi-supervised clustering while pairwise constraints are selected and updated in each iteration based on already existing clustering outcome in second category. Apparently, the initial selection is easy to operate with its weak-point (listed in last paragraph) being that some pairwise constraints are not properly selected. At last, this improper selection would result in poor clustering performance. As expected, the iterative selection could improve clustering performance since informative pairwise constraints are more likely to be selected while more computing time is needed. It could be considered that the iterative selection corresponds to active learning while the initial selection corresponds to non-active learning.

Zhong *et al.* [7] has listed two challenges that active learning faces. The first one is how to precisely generate pairwise constraints which have significant impact on clustering outcome. And the second is how to efficiently construct a series of query which could significantly reduce cost in the process of manual inspection to judge pairwise constraints.

In this paper, a novel active learning pairwise constraint formulation algorithm is proposed and it could overcome these two challenges successfully. Shannon entropy is used to depict and measure the uncertainty of samples while some of existing methods use neighborhood-based measure scheme [21]–[23]. Samples with great uncertainty are most likely involved in informative pairwise constraints. Hence these samples could be considered as informative data samples and we use them directly to construct a series of query in the new active learning algorithm proposed here. Some type of farthest-first strategy and priority principle are introduced with aim to enhance clustering performance and reduce cost in inspection.

The rest of this paper is organized as follows. Section II briefly review related work on active learning for semi-supervised clustering. In Section III, the new proposed active learning algorithm is introduced in details. In Section IV, empirical data experiment is conducted and the well-known MPCK-means (Metric Pairwise Constrained K-means) is used as the underlying constraint-based clustering algorithm. Experiment outcome shows the improvement of the proposed active learning algorithm in comparison with its counterparts. Some conclusion and future research work is presented in Section V.

II. RELATED WORK

Active learning has been extensively studied in supervised learning problems [11]–[13], [15]–[20], while in semi-supervised clustering its research is relatively limited [21].

Basu *et al.* [9] proposed an active query selection algorithm called Farthest First Query Selection (FFQS) algorithm. FFQS consists of two phases: Explore and Consolidate. In Explore phase, the farthest-first traversal method is proposed with aim to get disjoint non-null clusters (at least one point per cluster). In Consolidate phase, points which are not selected in Explore phase are picked randomly with aim to form query with points in each cluster obtained in Explore phase until a must-link constraint is confirmed and collected. Based on FFQS, Mallapragada *et al.* [22] proposed a modified version of FFQS called Min-Max and the min-max criterion is used to replace the random selection strategy in FFQS.

Xu *et al.* [24] proposed an active constrained spectral clustering algorithm by identifying boundary and sparse points based on spectral eigenvectors of the similarity matrix. The oracle (acting like a library set) is queried with aim to decide whether pairwise constraint is collected or not. Data experiment shows that this algorithm is creative and effective. However it is limited in the case of two clusters problem and error is considered to happen on only boundary points.

Huang and Lam [25] constructed an iterative framework with aim to discover pairwise constraints for semi-supervised text document clustering. In each iteration, the selection of pairwise constraints is related to the previous clustering result. Similar to Huang's method, Xiong *et al.* [21] proposed a neighborhood-based framework which focuses on the uncertainty of data points in terms of to which neighborhood it belongs rather than pairwise uncertainty. If a parallel querying system is available, these two methods probably could not use information effectively.

Zhong *et al.* [7] proposed a novel entropy-based active informative pairwise constraint formulation algorithm (AIPC) with aim to collect must-link constraints. AIPC consists of two phases: Pre-clustering and Marking. In Pre-clustering phase, some type of unsupervised clustering is used to get a preliminary membership degree matrix. In Marking phase, data sample are divided into two categories (strong and weak) in terms of sample uncertainty and a querying series is constructed and used to collect must-link constraints.

Cai *et al.* [23] proposed an active learning method which is a modified version of Min-Max algorithm. Both Explore and Consolidate phases act on an informative data set and the point with the greatest uncertainty is chosen as the first point in Explore phase. It is worth mentioning that the uncertainty measure used in Cai *et al.* [23] is depicted by point neighborhood defined therein.

Mainly inspired by the work in Zhong *et al.* [7] and Cai *et al.* [23], a novel active learning algorithm would be proposed in this paper. The entropy-based uncertainty measure will be used here in order to construct an informative data set and both cannot-link and must-link pairwise constraints are collected rather than only must-link type in Zhong *et al.* [7]. Both Explore and Consolidate phases [23] are adopted here and some type of new strategy and priority principle are introduced with aim to enhance clustering

TABLE 1. List of abbreviation.

Abbreviation	Explanation
FCM	Fuzzy c-means clustering
MPCK-means	Metric Pairwise Constrained K-means[4]
FFQS	Farthest First Query Selection[9]
Min-Max	An improved method of FFQS[22]
AIPC	Zhong's method[7]
$\mathcal{M}$	Set of must-link pairs
$\mathcal{C}$	Set of cannot-link pairs
NMI	Normalized Mutual Information[26]

performance and reduce cost in inspection. The time complexity of this new algorithm is $O(n)$ while that for Cai's method [23] is $O(n^2)$ (n is the data sample size).

III. ACTIVE LEARNING ALGORITHM

In this section, a new active learning pairwise constraint formulation algorithm based on skeleton sets (ALPCS) would be constructed with aim to formulate informative pairwise constraints efficiently and economically. The key point within ALPCS is how to construct a series of query between pairs of samples such that the number of queries is small as much as possible. Firstly, some related mathematical notations and concepts will be introduced in the following.

A. PRELIMINARY & PROBLEM FORMULATION

Assume the data sample set is $X \equiv \{x_1, \ldots, x_n\}$ containing n samples with x_j being the j-th sample. Denote the set of cannot-link pairs by $\mathcal{C}$ and the set of must-link pairs by $\mathcal{M}$. Obviously, $\mathcal{C}$ and $\mathcal{M}$ satisfy the following properties:

- $(x_j, x_k) \in \mathcal{M}$ & $(x_k, x_h) \in \mathcal{M} \Rightarrow (x_j, x_h) \in \mathcal{M}$
- $(x_j, x_k) \in \mathcal{M}$ & $(x_k, x_h) \in \mathcal{C} \Rightarrow (x_j, x_h) \in \mathcal{C}$

Besides similarity between samples, the type of pairwise (cannot-link & must-link) constraints is another important information source in semi-supervised clustering problem. And it could enhance the clustering accuracy rate. However, not all pairwise constraints play an important role in semi-supervised clustering problem. Once pairwise constraints are incorporated into some unsupervised clustering (eg. fuzzy c-means clustering, FCM) problem, some pairwise constraints may have a significant impact on clustering outcome while others have little impact. Zhong *et al.* [7] has presented detailed explanation on this phenomenon and divided pairwise constraints into two categories in terms of their impact on clustering outcome: **informative** and **non-informative** pairwise constraint. In general, **non-informative** pairwise constraints are considered to be invalid and redundant (even harmful) in semi-supervised clustering problem.

Since the collection of pairwise constraints may be rather time-consuming and costly, all active learning pairwise constraint algorithms including ALPCS proposed here are designed to collect **informative** pairwise constraints efficiently and economically.

B. METHODOLOGY

ALPCS consists of three phases: Selecting, Exploring and Consolidating. In Selecting phase, FCM is used to obtain an informative data set. In Exploring phase, the point with greatest uncertainty (Shannon entropy, **Definition 1**) in the informative data set is chosen as the first point, and then the farthest-first strategy (**Definition 2** & **Definition 3**) is used to construct clustering skeleton sets and collect pairwise constraints. If the number of skeleton sets equals the number of clusters, ALPCS gets into third phase Consolidating; otherwise, ALPCS would finish. In Consolidating phase, non-skeleton points included in the informative data set are used to construct a series of query with skeleton set representative points. The symmetric relative entropy minimum priority principle (**Definition 4** & **Definition 5**) is used to depict the similarity between two points within these queries. Figure 1 presents the flow chart of ALPCS.

1) SELECTING

Definition 1: Suppose that the data sample set X should be grouped into c clusters, and $\{\mu_{1j}, \cdots, \mu_{cj}\}$ *is the membership degree vector of* x_j*, for* $j \in \{1, \cdots, n\}$*. This means that* μ_{ij} *is the probability of* x_j *belonging to the i-th cluster. Define the Shannon entropy for* x_j *by*

$$E(x_j) = -\sum_{i=1}^{c} \mu_{ij} \ln \mu_{ij}. \tag{1}$$

In general, Shannon entropy is used to depict uncertainty degree of sample point. When $\mu_{ij} = \frac{1}{c}$ *for all* $i \in \{1, \cdots, c\}$*, Shannon entropy reaches the maximum*

$$Max(E(x_j)) = \ln c.$$

At this time, uncertainty gets its maximum. Generally speaking, the greater Shannon entropy of sample point, the greater uncertainty.

In Selecting phase, FCM algorithm is used to generate an informative data set by selecting samples with greater uncertainty. The objective function of FCM is defined by

$$\begin{aligned} J_{FCM}(U, V) &= \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{m} ||x_j - v_i||^2 \\ s.t. \quad & 0 \le \mu_{ij} \le 1, \quad \sum_{i=1}^{c} \mu_{ij} = 1 \\ & 1 \le i \le c, \quad 1 \le j \le n, \end{aligned} \tag{2}$$

where $m(m > 1)$ is the degree of fuzziness, $||x_j - v_i||$ represents the Euclidean distance between x_j and v_i, $U \equiv [\mu_{ij}]$ is the membership degree matrix and $V \equiv [v_1, v_2, \cdots, v_c]$ consists of c center v_is of clusters. Using Lagrange multiplier method to minimize J_{FCM}, two iterative equations are obtained as follows

$$v_i = \frac{\sum_{j=1}^{n} x_j \mu_{ij}^{m}}{\sum_{j=1}^{n} \mu_{ij}^{m}}, \tag{3}$$

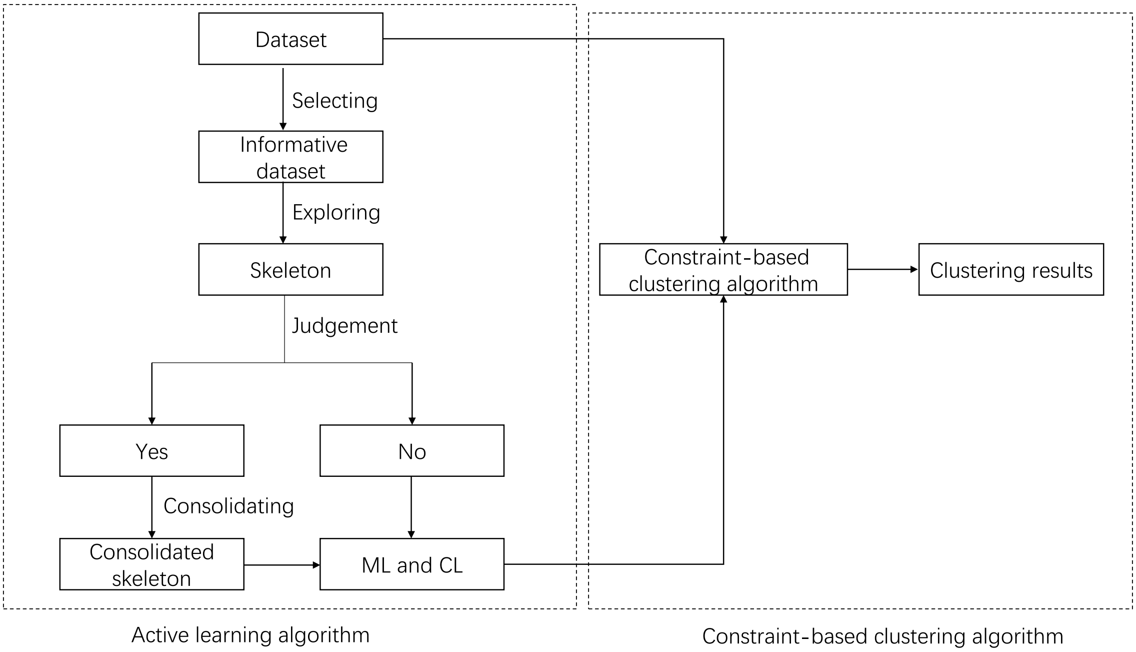


FIGURE 1. Flow chart of ALPCS.

and

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}(\frac{||x_j - v_i||}{||x_j - v_k||})^{\frac{2}{m-1}}}. \tag{4}$$

Iteration in FCM will terminate when $|J_{FCM}{}^{(t+1)} - J_{FCM}{}^{(t)}| \leq \epsilon_0$ (admissible error) or t (number of cycles) reaches the maximum number T of iteration given in advance.

Based on the ultimate membership degree matrix in FCM, Shannon entropy for all sample points are calculated. Then sort all sample points in decreasing order in terms of Shannon entropy, and select the top $p \times 100\%$ $(0 < p \leq 1)$ part of them as the informative data set (S_1 in **Algorithm 1**). Both in Exploring phase and Consolidating phase, this informative data set is a starting point.

2) EXPLORING

In Exploring phase, some type of distance (**Definition 2**) between point and set is needed.

Definition 2: Assume a point x and a set $Y = \{y_1, \ldots, y_w\}$. Define the distance between x and Y by

$$d(x, Y) = \frac{\sum_{j=1}^{w} ||y_j - x||}{w}, \tag{5}$$

where w is the element number (cardinality) of set Y and $||y_j - x||$ is the Euclidean distance between y_j and x.

Based on **Definition 2**, Farthest-first strategy is proposed with aim to update skeleton sets in each iteration of **Algorithm 2**.

Definition 3: Farthest-first strategy is that the point (from surplus informative data set) which is farthest away from already existing skeleton sets is the preferred choice treated as one point of query series in next iteration.

Algorithm 1 Selecting

```
Input: Data sample set $X$; the number of clusters $c$; the maximum number of iteration $T$; the ratio $p$ of informative sample points;
Output: Informative data set $S_1$;
Initialization: set $U = [\mu_{ij}^{(0)}]$ (some initial membership degree matrix), the iteration number $t = 0$;
repeat
    Update cluster center $v_i$s by (3);
    Update membership degree $\mu_{ij}$s by (4);
    Updata iteration number $t \leftarrow t + 1$;
until $|J_{FCM}^{(t+1)} - J_{FCM}^{(t)}| \leq \epsilon_0$ or $t = T$
Calculate Shannon entropy for all sample points and sort them in terms of Shannon entropy in decreasing order;
Select the top $p \times 100$ part of all sample points as the informative data set $S_1$;
return $S_1$
```

Based on the informative data set S_1, a skeleton set structure would be constructed in Exploring phase. In this process, element from S_1 is chosen one by one as one point of query series, the informative data set shrinks and skeleton set structure grows by degrees. Since the farthest point is most likely to succeed in constructing **informative** pairwise constraint, Farthest-first strategy is adopted.

In the process of constructing skeleton set structure, the point with the greatest uncertainty in S_1 is picked up as the first (initial) point while FFQS [9] selects the first point randomly. This manipulation could reduce randomness and it is reasonable to choose the point with the greatest uncertainty as the first point since the greatest uncertainty most likely corresponds to the type of **informative** pairwise constraints. Further, by Farthest-first strategy, the point (in surplus informative data set) which is farthest away from already existing skeleton sets is selected and used to design a series of query with all points in already existing skeleton sets. If these queries are all cannot-link, a new skeleton set would be constructed and this farthest point is its unique element. Otherwise, there must be at least one of these queries which is must-link. Then this farthest point is incorporated into some existing skeleton set to which this must-link query corresponds. This process continues until iteration runs out of its upper limit or c skeleton sets are already constructed.

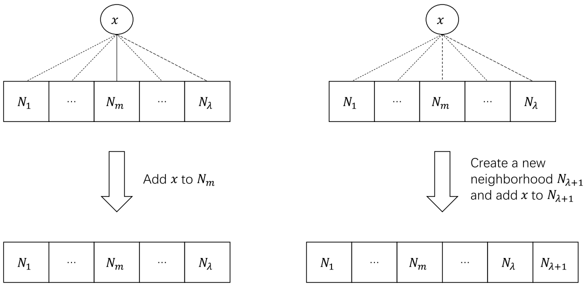


FIGURE 2. Overview of exploring phase. The solid line indicates must-link constraint and the dotted line indicates cannot-link constraint.

The key procedure in **Algorithm 2** is shown in Figure 2. Pairwise constraint sets ($\mathcal{C}$ and $\mathcal{M}$ outputted by **Algorithm 2**) are exactly what the constraint-based clustering algorithm needs in next clustering stage, refer to Figure 1.

```
Algorithm 2 Exploring
Input: Informative data set $S_1$; the number of clusters $c$; the maximum number of queries $Q$;
Output: $\lambda(\lambda \leq c)$ disjoint skeleton set $\{N_t\}_{t=1}^{\lambda}$ with at least one point per set; cannot-link constraint set $\mathcal{C}$; must-link constraint set $\mathcal{M}$; surplus informative data set $S_2$;
Initialization: set $\{N_t\}_{t=1}^{\lambda}$, $\mathcal{C}$ and $\mathcal{M}$ to null, the iteration number $q = 0$;
Pick the point $x$ with the greatest uncertainty in $S_1$, set $\lambda \leftarrow 1$ and $N_1 \leftarrow \{x\}$, update $S_1 \leftarrow S_1 - \{x\}$;
repeat
    Pick the point $x$ $(\in S_1)$ which is farthest away from already existing skeleton set $\{N_t\}_{t=1}^{\lambda}$;
    Construct a series of query between $x$ and all points in already existing skeleton sets;
    if these queries are all cannot-link then
        Update $\lambda \leftarrow \lambda + 1$, construct a new skeleton set $N_\lambda = \{x\}$, and add all these cannot-link constraints into $\mathcal{C}$;
    else
        Pick one must-link constraint (recommend the first one in decreasing uncertainty searching order) and add $x$ into the existing skeleton set to which this must-link query corresponds; Add this must-link constraint into $\mathcal{M}$;
    end if
    Update $S_1 \leftarrow S_1 - \{x\}$;
    Update iteration number $q \leftarrow q + 1$;
until Obtain $c$ disjoint skeleton sets with at least one point per set or $q = Q$
$S_2 = S_1$;
return $\{N_t\}_{t=1}^{\lambda}$, $\mathcal{C}$, $\mathcal{M}$, $S_2$
```

3) CONSOLIDATING

If the number of skeleton sets obtained in Exploring phase equals the cluster number c, ALPCS gets into third phase Consolidating; otherwise, ALPCS would finish in second phase. Since surplus informative data set S_2 outputted by **Algorithm 2** maybe include significant information for **informative** pairwise constraint, third phase Consolidating is really needed to dig must-link constraints included in it.

In this phase, a difference measure between membership vectors of two sample points in each query is needed which is defined in **Definition 4**.

Definition 4: Define relative entropy between two sample point x_j and x_k by

$$D_{KL}(x_j||x_k) = \sum_{i=1}^{c} \mu_{ij} \ln \frac{\mu_{ij}}{\mu_{ik}}, \quad (1 \leq j, k \leq n),$$

where μ_{ij} is the membership degree of the j-th sample belonging to the i-th cluster and μ_{ik} is the membership degree of the k-th sample belonging to the i-th cluster. Relative entropy is also called Kullback-Leibler divergence which is used to depict the difference between two probability distributions. The greater $D_{KL}(x_j||x_k)$, the greater difference between x_j and x_k. Considering the asymmetry property of relative entropy, symmetric relative entropy is introduced and defined by

$$D_{SKL}(x_j||x_k) = \frac{1}{2}\left(D_{KL}(x_j||x_k) + D_{KL}(x_k||x_j)\right). \qquad (6)$$

The greater $D_{SKL}(x_j||x_k)$, the greater difference between x_j and x_k.

Based on **Definition 4**, a priority principle is introduced as follows.

Definition 5: The symmetric relative entropy minimum priority principle is that when a non-skeleton set point x in surplus informative data set S_2 is judged whether to be added into a skeleton set or not, y in skeleton sets with minimum symmetric relative entropy $D_{SKL}(x||y)$ is the preferred choice to form a query with x.

Since the number of skeleton sets is c in Consolidating phase, there are at most $c - 1$ times needed to determine to which skeleton set the non-skeleton point from S_2 belongs. According to the symmetric relative entropy minimum priority principle, we prefer to choose the point u_t that minimizes the value of $D_{SKL}(x, u_t)$ to form a query with x. In each iteration, the point x with the greatest uncertainty in updated surplus informative data set S_2 is selected, and then representative point u_ts in skeleton set $\{N_t\}_{t=1}^{c}$ which are closest to x are selected. The technique of choosing representative points for skeleton sets is motivated by [9], and it could reduce the number of queries obviously and significantly.

Algorithm 3 Consolidating

```
Input: Surplus informative data set $S_2$; the number of clusters $c$; the maximum number of queries $Q$; $c$ disjoint skeleton set $\{N_t\}_{t=1}^{c}$ with at least one point per set; must-link constraint set $\mathcal{M}$;
Output: Updated must-link constraint set $\mathcal{M}$;
Set the iteration number $q = 0$;
repeat
    Pick the point $x$ ($\in S_2$) with the greatest uncertainty;
    Select representative point $u_t$s for all skeleton set $\{N_t\}_{t=1}^{c}$ such that $u_t$ is closest to $x$;
    Calculate the symmetric relative entropy $D_{SKL}(x, u_t)$ between $x$ and $u_t$ ($\in N_t$);
    Rearrange $D_{SKL}(x, u_t)$ in increasing order and denote this sorted entropy series by $\{D_{SKL}(x, u_{t_{(1)}}), \cdots, D_{SKL}(x, u_{t_{(c)}})\}$;
    for $h = 1$ to $c$ do
        Seek answer to the query $(x, u_{t_{(h)}})$ till must-link is obtained; and then add this must-link constraint into $\mathcal{M}$; add $x$ into the skeleton set to which this must-link query corresponds;
    end for
    Update $S_2 \leftarrow S_2 - \{x\}$;
    Update iteration number $q \leftarrow q + 1$;
until $q = Q$ or $S_2$ turns into a null set;
return $\mathcal{M}$
```

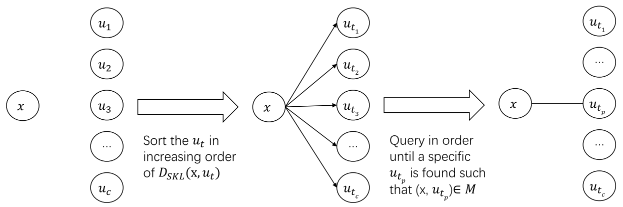

x
u_1
u_2
u_3
...
u_c
Sort the u_t in increasing order of $D_{SKL}(x, u_t)$
x
u_{t_1}
u_{t_2}
u_{t_3}
...
u_{t_c}
Query in order until a specific u_{t_p} is found such that (x, u_{t_p})∈ M
x
u_{t_1}
...
u_{t_p}
...
u_{t_c}

FIGURE 3. Overview of consolidating phase. The arrow indicates the query and the solid lines indicates must-link constraint.

The key procedure in **Algorithm 3** is shown in Figure 3 and the must-link constraint set $\mathcal{M}$ from **Algorithm 2** is updated by **Algorithm 3**.

IV. EXPERIMENTS

In this section, the well-known MPCK-means (Metric Pairwise Constrained K-means) semi-supervised clustering algorithm [4] is used as the underlying constraint-based clustering algorithm. The performance of ALPCS is evaluated in comparison with its related counterparts in six empirical dataset experiments.

A. DATASETS

Six benchmark UCI datasets (Iris, Wine, Breast, Heart, Parkinsons and Ecoli) have been extensively analyzed with aim to evaluate the performance of constraint-based clustering algorithms [21], [23]. We also use them to do experiments in this section. For dataset Ecoli, the smallest three classes which only contain 2, 2 and 5 instances respectively are deleted in advance. Table 2 lists characteristics of these six datasets.

TABLE 2. Characteristics of datasets.

Datasets	# of Samples	# of Features	# of Clusters
Iris	150	4	3
Wine	178	13	3
Breast	683	9	2
Heart	270	12	2
Parkinsons	195	22	2
Ecoli	327	7	5

B. COMPARATIVE METHOD

In order to demonstrate the effectiveness of ALPCS proposed in this paper, five counterparts are considered here: Random policy, FFQS [9], Min-Max method [22], Cai's method [23] and AIPC [7].

Random policy: points are totally randomly selected in order to form pairwise constraints and this policy is commonly used as comparative baseline in active learning study.

FFQS: two phases (Explore and Consolidate) are included. In Explore phase, some type of farthest-first strategy is used to construct disjoint neighborhoods, at least one point per neighborhood. In Consolidate phase, non-neighborhood points are selected randomly to form queries with each point in neighborhoods until a must-link pair is obtained.

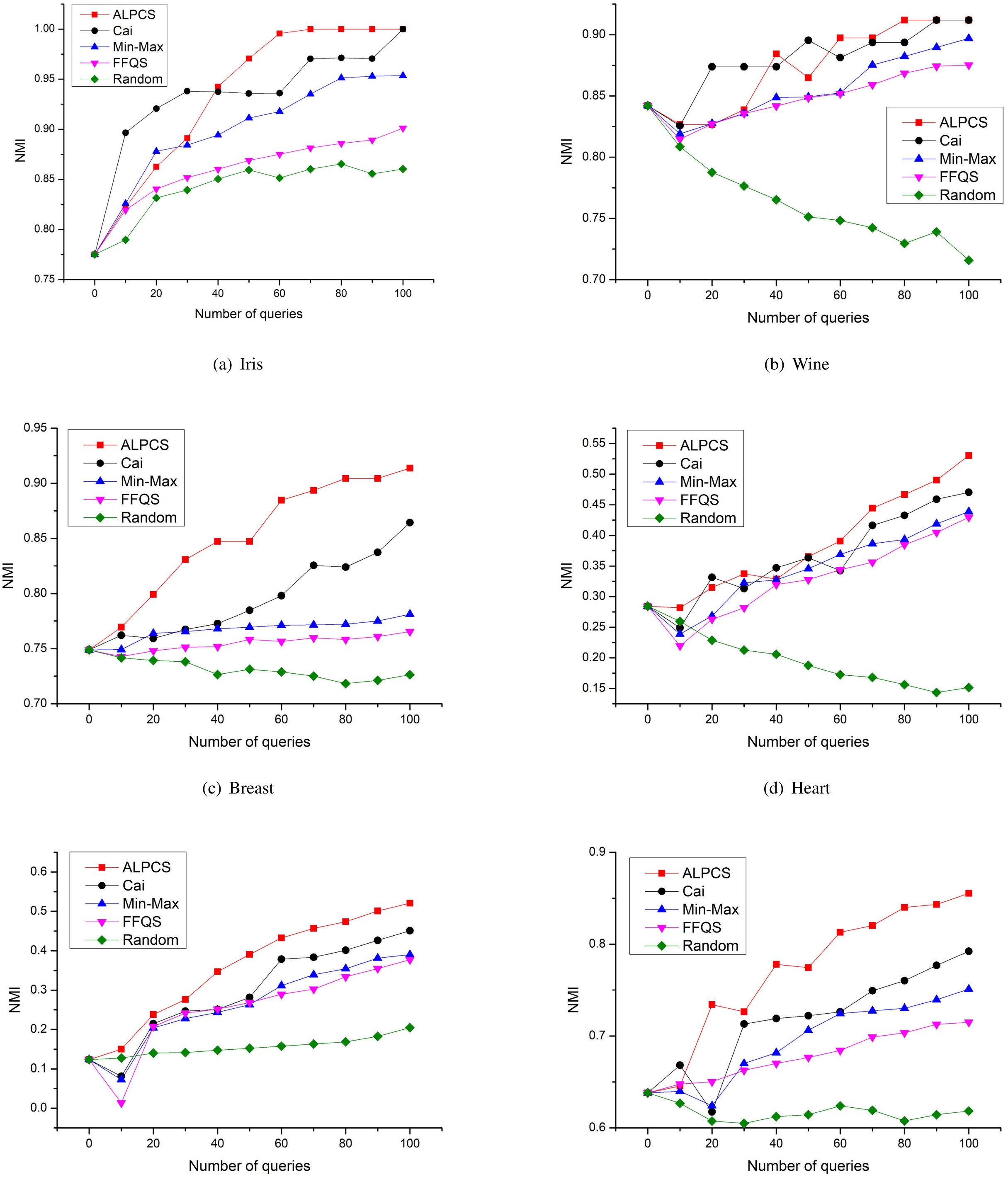
Min-Max: this method is a modified version of FFQS. A type of Min-Max criterion is proposed with aim to replace the random selection strategy in FFQS.

Cai's method: this method is a modified version of Min-Max. A notation of information data set is introduced and point selection strategy is based on an information data set.

AIPC: samples are divided into two categories (weak and strong) in terms of entropy-based uncertainty. A type of priority principle is introduced and used to construct queries between weak and strong points.

C. TIME COMPLEXITY ANALYSIS

ALPCS consists of three phases. The first phase is similar to Pre-clustering phase in Zhong *et al.* [7] and the remaining two phases are similar two phases in Basu *et al.* [9]. Hence the time complexity of ALPCS is a compound body of FFQS and AIPC. Table 3 lists the time complexity of all algorithms involved in this paper. The time complexity for Cai's method [23] is $O(n^2)$.



(a) Iris

(b) Wine

(c) Breast

(d) Heart

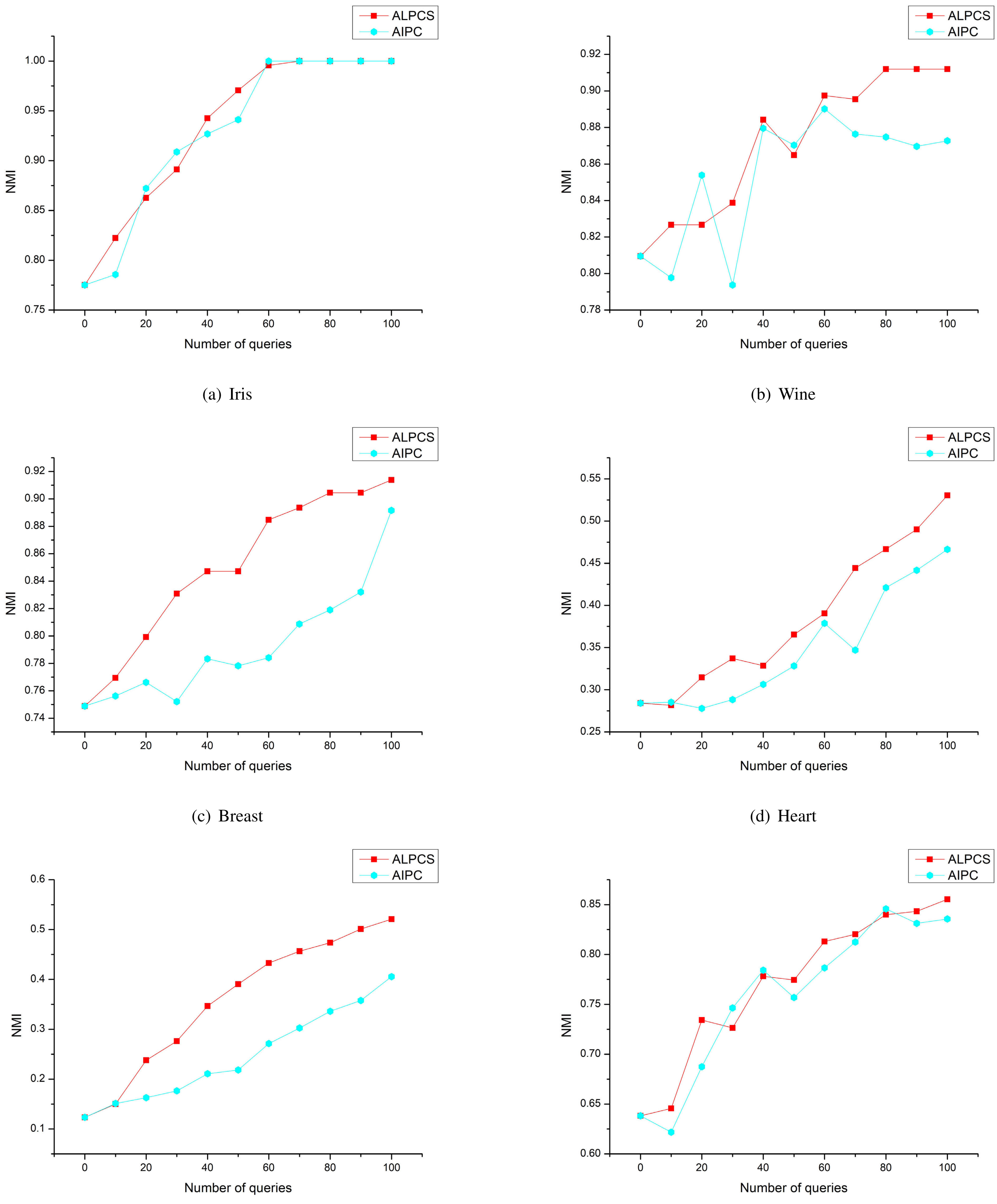
(e) Parkinsons

(f) Ecoli

FIGURE 4. NMI values of different methods on six datasets as a function of the number of pairwise queries.

D. EVALUATION CRITERIA

The index NMI (Normalized Mutual Information) is used to evaluate the clustering assignments against the ground truth class labels [26]. NMI considers both the actual class label and the predicted clustering assignment as random variables and measures the mutual information between these two random variables. It is normalized to a zero-to-one range. If C is the random variable representing the cluster assignments of



(a) Iris

(b) Wine

(c) Breast

(d) Heart

(e) Parkinsons

(f) Ecoli

FIGURE 5. NMI values of different methods on six datasets as a function of the number of pairwise queries.

instances and K is the random variable representing the class labels of instances, then NMI is defined by

$$NMI = \frac{2I(C;K)}{H(C)+H(K)},$$

where $I(C;K) = H(C) - H(C|K)$ is the mutual information between these two random variables, $H(C)$ is the entropy of C and $H(C|K)$ is the the conditional entropy of C given K. The closer value of NMI to 1, the better clustering performance.

TABLE 3. The time complexity.

Method	Time complexity
ALPCS	$O(n)$
AIPC	$O(n)$
Cai	$O(n^2)$
Min-Max	$O(n)$
FFQS	$O(n)$

E. EXPERIMENTAL SETTING

The well-known MPCK-means semi-supervised clustering algorithm [4] is used as the underlying constraint-based clustering algorithm after all active learning algorithms involved here have been executed. The iteration number of MPCK-means is set to be 100 and other parameters within it use default values.

In each experiment, around 100 pairwise queries are collected. The answer for queries is verified by inspection according to the true class label in each dateset. Then MPCK-means is used to execute clustering based on pairwise constraints obtained in active learning stage. In order to reduce randomness of both active learning and MPCK-means, 50 independent repetitions are used to estimate NMI.

The parameter p values in Selecting phase of ALPCS corresponding to datasets analysed here are listed in Table 4. n_1 (the number of points in the informative data set, this is $n_1 = p \times$ sample size of data set) is chosen such that the total number of queries obtained in Exploring phase reaches around 100.

TABLE 4. Values of p given in advance.

Datesets	# of Samples	n_1	p
Iris	150	103	0.6867
Wine	178	136	0.7640
Breast	683	103	0.1508
Heart	270	102	0.3778
Parkinsons	195	102	0.5231
Ecoli	327	108	0.3303

For n_1, in our experiment, we set the maximum number of queries to be 100, so n_1 should satisfy two conditions: 1) it can get 100 queries; 2) it is the minimum value that satisfies condition 1. We use a simple iterative method to calculate n_1, let $n_1 = n - i(i = 1, 2, \ldots, \text{n})$, when we get a special i_1 that makes n_1 satisfy condition 1, at the same time $i_1 - 1$ makes n_1 not satisfy condition 1, then obviously i_1 also makes n_1 satisfy condition 2. Finally, after obtaining the definite n_1, we also get the corresponding p.

F. EXPERIMENTAL RESULT

Figure 4 and 5 show the result of experiments: the x-axis is the number of pairwise queries and the y-axis is the value of NMI. Each curve denotes the mean of 50 repetitions independently. The more pairwise constrains selected by active learning algorithms except the random policy, the better clustering performance of MPCK-means. Counterparts in Figure 4 are four neighborhood-based algorithms (Cai,Min-Max,FFQS and Random policy) and counterpart in Figure 5 is a recent entropy-based method (AIPC).

In Figure 4, for Breast, Parkinsons and Ecoli, the performance of ALPCS improves significantly with the increase of the number of pair constraints and is always better than the other four counterparts. For Iris and Heart, although ALPCS is not as good as its counterparts under a small number of pairwise constraints, it shows a better performance under a large number of pairwise constraints. For Wine, ALPCS has some fluctuations and is still competitive. Fig. 5 shows that ALPCS is comparable to AIPC for Iris, Wine and Ecoli and for Breast, Heart and Parkinsons, ALPCS has better performance. In a word, ALPCS provides a better (at least competitive) selection of pairwise constraints in the clustering process.

It is confirmed that the random policy degrades the clustering performance dramatically as the number of pairwise constraints increases in four datasets (Breast, Heart, Wine and Ecoli). This phenomenon has been shown in previous study and demonstrates the importance of pairwise constraint proper selection. Of three counterparts (FFQS, Min-Max, Cai's method), the performance of Cai's method is the best and FFQS is the worst. This is consistent with the relationship of these three algorithms.

V. CONCLUSION AND FUTURE WORK

In this paper, a new active learning method (ALPCS) is proposed for semi-supervised clustering. A Shannon entropy-based uncertainty measure is used here to construct an informative data set. And some type of strategy and priority principle are introduced to construct queries between points in the informative data set and collect two types of pairwise constraints with aim to enhance clustering performance and reduce cost in manual inspection. Data experiment shows that ALPCS provides a better (at least competitive) selection of pairwise constraints in the clustering process.

ALPCS is mainly inspired by the work in Zhong *et al.* [7] and Cai *et al.* [23]. Both cannot-link and must-link pairwise constraints are collected by ALPCS rather than only must-link type by AIPC in Zhong *et al.* [7]. Further, the time complexity of ALPCS is $O(n)$ while that for Cai's method [23] is $O(n^2)$.

In Selecting phase of ALPCS, FCM is recommended here since its good performance in depicting uncertainty of samples. There have been some even better alternatives [27]–[31] to FCM which could be used here. Besides, cross entropy [32], [33] could be used here to replace relative entropy in Consolidating phase. However, this requires great difference between different classes otherwise it would increase cost in manual inspection. The clustering performance would be influenced by unbalance and dimension of data set significantly. As a reviewer pointed out that unbalance due to scarcity of samples in some categories could be alleviated by over-sampling and subrogation methods or simply using replicates of the original data. These would be research directions for modified version of ALPCS in the future.

REFERENCES

[1] P. Balakrishnan, M. Cooper, V. Jacob, and P. Lewis, "A study of the classification capabilities of neural networks using unsupervised learning: A comparison with$K$-means clustering," *Psychometrika*, vol. 59, no. 4, pp. 509–525, 1994.

[2] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *Proc. 18th Int. Conf. Mach. Learn.*, San Francisco, CA, USA, 2001, pp. 577–584.

[3] K. Wagstaff and C. Cardie, "Clustering with instance-level constraints," in *Proc. 17th Int. Conf. Mach. Learn.*, San Francisco, CA, USA, 2000, pp. 1103–1110.

[4] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, Banff, ON, Canada, 2004, pp. 81–88.

[5] Z. Yu, P. Luo, J. Liu, H.-S. Wong, J. You, G. Han, and J. Zhang, "Semi-supervised ensemble clustering based on selected constraint projection," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 12, pp. 2394–2407, Dec. 2018.

[6] J. Zhou, Z. Lai, C. Gao, X. Yue, and W. Wong, "Rough-fuzzy clustering based on two-stage three-way approximations," *IEEE Access*, vol. 6, pp. 27541–27554, 2018.

[7] G. Zhong, X. Deng, and S. Xu, "Active informative pairwise constraint formulation algorithm for constraint-based clustering," *IEEE Access*, vol. 7, pp. 81983–81993, 2019.

[8] Q. Lei and T. Li, "Semi-supervised selective affinity propagation ensemble clustering with active constraints," *IEEE Access*, vol. 8, pp. 46255–46266, 2020.

[9] S. Basu, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering," in *Proc. SIAM Int. Conf. Data Mining*, Lake Buena Vista, FL, USA, Apr. 2004, pp. 333–344.

[10] D. Greene and P. Cunningham, "Constraint selection by committee: An ensemble approach to identifying informative constraints for semi-supervised clustering," in *Proc. Eur. Conf. Mach. Learn.* Warsaw, Poland, 2007, pp. 140–151.

[11] Y. Liu, K. Liu, C. Zhang, X. Wang, S. Wang, and Z. Xiao, "Entropy-based active sparse subspace clustering," *Multimedia Tools Appl.*, vol. 77, no. 17, pp. 22281–22297, Sep. 2018.

[12] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Batch mode active learning and its application to medical image classification," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA, 2006, pp. 417–424.

[13] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *Int. J. Comput. Vis.*, vol. 113, no. 2, pp. 113–127, Jun. 2015.

[14] J. Wu, V. S. Sheng, J. Zhang, H. Li, T. Dadakova, C. L. Swisher, Z. Cui, and P. Zhao, "Multi-label active learning algorithms for image classification: Overview and future promise," *ACM Comput. Surv.*, vol. 53, no. 2, pp. 1–35, 2020.

[15] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA, 2004, pp. 623–630.

[16] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, no. 1, pp. 999–1006, 2002.

[17] S. C. H. Hoi, R. Jin, and M. R. Lyu, "Large-scale text categorization by batch mode active learning," in *Proc. 15th Int. Conf. World Wide Web (WWW)*, New York, NY, USA, 2006, pp. 633–642.

[18] A. Bartoli, A. De Lorenzo, E. Medvet, and F. Tarlao, "Active learning of regular expressions for entity extraction," *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 1067–1080, Mar. 2018.

[19] A. Shelmanov, V. Liventsev, D. Kireev, N. Khromov, A. Panchenko, I. Fedulova, and D. V. Dylov, "Active learning with deep pre-trained models for sequence tagging of clinical and biomedical texts," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, San Diego, CA, USA, Nov. 2019, pp. 482–489.

[20] A. Bartoli, A. De Lorenzo, E. Medvet, and F. Tarlao, "Active learning approaches for learning regular expressions with genetic programming," in *Proc. 31st Annu. ACM Symp. Appl. Comput. (SAC)*, New York, NY, USA, 2016, pp. 97–102.

[21] S. Xiong, J. Azimi, and X. Z. Fern, "Active learning of constraints for semi-supervised clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 43–54, Jan. 2014.

[22] P. K. Mallapragada, R. Jin, and A. K. Jain, "Active query selection for semi-supervised clustering," in *Proc. 19th Int. Conf. Pattern Recognit.*, Tampa, FL, USA, Dec. 2008, pp. 1–4.

[23] L. Cai, T. Yu, T. He, L. Chen, and M. Lin, "Active learning method for constraint-based clustering algorithms," in *Proc. Int. Conf. Web-Age Inf. Manage.*, Nanchang, China, 2016, pp. 319–329.

[24] Q. Xu, M. desJardins, and K. Wagstaff, "Active constrained clustering by examining spectral eigenvectors," in *Discovery Science*. Berlin, Germany: Springer-Verlag, 2005, pp. 294–307.

[25] R. Huang and W. Lam, "Semi-supervised document clustering via active learning with pairwise constraints," in *Proc. 7th IEEE Int. Conf. Data Mining (ICDM)*, Las Vegas, NV, USA, Oct. 2007, pp. 517–522.

[26] L. I. Kuncheva and S. T. Hadjitodorov, "Using diversity in cluster ensembles," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, The Hague, The Netherlands, Oct. 2004, pp. 1214–1219.

[27] D.-Q. Zhang and S.-C. Chen, "A novel kernelized fuzzy C-means algorithm with application in medical image segmentation," *Artif. Intell. Med.*, vol. 32, no. 1, pp. 37–50, Sep. 2004.

[28] S. Wang, K. F. L. Chung, Z. Deng, D. Hu, and X. Wu, "Robust maximum entropy clustering algorithm with its labeling for outliers," *Soft Comput.*, vol. 10, no. 7, pp. 555–563, 2006.

[29] L. Zhu, F.-L. Chung, and S. Wang, "Generalized fuzzy C-Means clustering algorithm with improved fuzzy partitions," *IEEE Trans. Syst., Man, Cybern., B (Cybern.)*, vol. 39, no. 3, pp. 578–591, Jun. 2009.

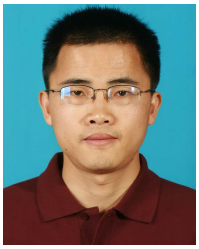
[30] S. R. Kannan, S. Ramathilagam, P. Devi, and A. Sathya, "Improved fuzzy clustering algorithms in segmentation of DC-enhanced breast MRI," *J. Med. Syst.*, vol. 36, no. 1, pp. 321–333, Feb. 2012.

[31] G. Tseng, "Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data," *Bioinformatics*, vol. 23, no. 17, pp. 2247–2255, 2007.

[32] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, Feb. 2005.

[33] J. Ye, "Single valued neutrosophic cross-entropy for multicriteria decision making problems," *Appl. Math. Model.*, vol. 38, no. 3, pp. 1170–1175, Feb. 2014.

DUO WEN CHEN received the B.S. degree from Inner Mongolia University, Hohhot, China, in 2016. He is currently pursuing the M.S. degree with the Guangdong University of Technology, Guangzhou, China. His research interests include data clustering and computational intelligence.



YING HUA JIN received the B.S. degree in mathematics from Southwest University, Chongqing, China, in 2004, and the M.S. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2009. In 2011, he joined as a Staff of the Guangdong University of Technology, where he is currently a Lecturer with the School of Apply Mathematics. His research interests include log-linear model and data clustering.