IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 6, 2019, accepted June 14, 2019, date of publication June 19, 2019, date of current version July 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923659

Active Informative Pairwise Constraint Formulation Algorithm for Constraint-Based Clustering

GUOXIANG ZHONG[1], XIUQIN DENG[1], AND SHENGBING XU[1,2]
[1]School of Apply Mathematics, Guangdong University of Technology, Guangzhou 510520, China
[2]School of Computer, Guangdong University of Technology, Guangzhou 510006, China

Corresponding author: Shengbing Xu (xushengbing111@126.com)

This work was supported in part by the Natural Science Foundation of China under Grant 61672169, and in part by the Soft Science Research Project in Guangdong Province under Grant 2015A070704049.

ABSTRACT Constraint-based clustering utilizes pairwise constraints to improve clustering performance. In this paper, we propose a novel formulation algorithm to generate more informative pairwise constraints from limited queries for the constraint-based clustering. Our method consists of two phases: pre-clustering and marking. The pre-clustering phase introduces the fuzzy c-means clustering (FCM) to generate the cluster knowledge that is composed of the membership degree and the cluster centers. In the marking phase, we first propose the weak sample with the larger uncertainty expressed by the entropy of the membership degree. Then, we study the strong sample that contains less uncertainty and should be closest to its cluster center. Finally, given weak samples in descending order of entropy, we formulate informative pairs with strong samples and seek answers using the second minimal symmetric relative entropy priority principle, which leads to more efficient queries. Making use of the pairwise constraint k-means clustering (PCKM) as the underlying constraint-based clustering algorithm, further data experiments are conducted in several datasets to verify the improvement of our method.

INDEX TERMS Constraint-based clustering, pairwise constraint, weak sample, strong sample, symmetric relative entropy.

I. INTRODUCTION

The constraint-based clustering which applies prior information in the form of pairwise constraint has more improvement in clustering performance [1]–[6]. The pairwise constraint defines the cluster relationship between two samples [7]. The types of it include must-link pair and cannot-link pair which indicate two samples are in same clusters or not respectively. In general, the user of constraint-based clustering would first formulate a list of pairs and then attains the answers through active learning. With the help of these pairwise constraints, constraint-based clustering takes advantage of attributes to discover the underlying clustering standard between samples and groups a set of samples into fixed number clusters. It is noteworthy that two categories of general pairwise constraint formulation frameworks are explored to enforce the clustering result in the existing research studies.

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

The first category of approaches methods [8], [9] alleviate the challenges in query view which obtain must-link pair from limited queries as more as possible. The second category of approaches methods based on sample view include boundary-sample-based method [10], neighborhood-sample-based schemes [11], [12] and informative-sample-based framework [13] and so on.

But despite their apparent success, the problem has not been fundamentally improved in pair view. This view does not conflict but consolidate with two previous views. It would confirm more must-link pairs and construct pairwise constraints with samples which would impact the clustering performance. Two main challenges they suffer from is: 1) how to generate informative pairwise constraint whose efficiency of improving clustering performance is more powerful; 2) how to make queries more efficient for the user practical application goal of cutting cost-consuming.

The answer of pairs considered by those above-mentioned formulation algorithms are provided by active learning

which is mostly applied in labeling and sampling of image processing [14]–[17], natural language processing [18], [19] and so on [20], [21]. According to sampling strategy, active learning can be divided into three categories: uncertainty sampling [14]–[17], [22], [23], query-by-committee [24]–[27], expected error minimization [28]–[31]. The uncertainty sampling is one of the most common methods due to its simplicity. By its definition, this method would be designed to select samples which have less certainty. Yang *et al.* discussed the uncertainty sampling with taking active multi-class scenario into account [14], in which the measure of certainty of sample was active pool and diversity information was added into objective function. As discussed in [15], some multi-class active learning did not consider the uncertainty of outliers. Du *et al.* proposed a robust multi-label active learning algorithm which introduced maximum correntropy criterion as the measure of uncertainty [15]. In [16], [17], the sparse modeling was incorporated into this samples selection to address the problem of redundant information between uncertain samples. In query-by-committee method, the query is formulated according to the criterion of minimal agreement. And it attains more and more information about classifier. But this may lead to ignore the difference between different committees. The expected error minimization method selects the unlabeled sample which minimize the expected error of the current classifier. On account of its computational expense, this method always be used in binary classifier. As for constraint-based clustering, active learning was introduced by Basu in 2004 [8]. Besides, the International Journal of Computer Vision (IJCV), one of computer vision top journal, also had published a special issue on active learning in 2013. The reason which many scholars are interested in the research of it is that the labels are no longer quite so expensive and time-consuming but the performance of models are still powerful [32]–[34].

In this paper, we propose active informative pairwise constraint formulation algorithm (AIPC) to solve the above-mentioned problem. Table 1 provides the comparison of various formulating pairwise constraint methodologies. The AIPC has two phases: Pre-clustering and Marking. In the Pre-clustering phase, fuzzy c-means clustering (FCM) is introduced to develop cluster knowledge that contains membership degree and cluster centers. The membership degree provides the fuzzy type of cluster label to measure the uncertain belonging of each sample. The cluster center is the mean of all samples in the same cluster, weighted by their membership degree. It contains less uncertainty and represents the clustering pattern of each cluster. In the Marking phase, we study the membership degree to present an entropy approach for measuring the uncertainty associated with each sample. The larger the entropy of the sample, the larger is the uncertainty. If the entropy is greater than one threshold value, the sample is referred to as weak sample. Then we should select strong sample which is closest to its cluster center always has less uncertainty in clustering. Given the weak sample in descending order of entropy, the queries formed with strong samples use the second minimal symmetric relative entropy priority principle until a must-link pair is obtained. Weak samples and strong samples make up informative pairwise constraints that can improve the clustering performance. The second minimal symmetric relative entropy priority principle makes the query more efficient.

TABLE 1. The comparison of various formulating pairwise constraint methodologies.

Method	1^{st} category	2^{nd} category	Proposed
View	Query	Sample	Pair

The contributions of the paper are twofold. First, we propose informative pairwise constraint that include weak sample and strong sample which contain less uncertainty and more uncertainty respectively. Second, following the second minimal symmetric relative entropy priority principle, obtaining the answer of the query leads to lower cost.

The rest of this paper is organized as follows. In Section II, we introduce our proposed pairwise constraint algorithm AIPC in detail. There are three cases, preliminary, problem formulation and methodology. Then, in Section III, we provide the data experiments conducted in different datasets and the underlying constraint-based clustering algorithm PCKM. The experimental results verify the improvement of AIPC over comparative methods. Finally, some conclusions and future work are presented in Section IV.

II. ACTIVE INFORMATIVE PAIRWISE CONSTRAINT ALGORITHM

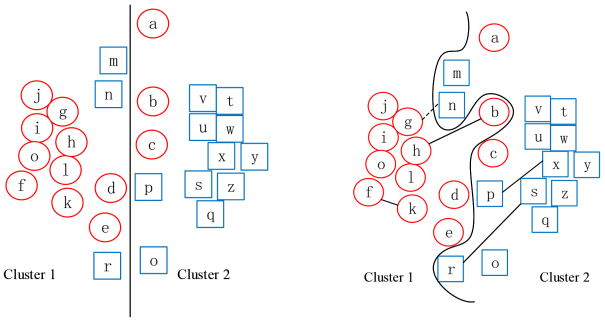
A. PRELIMINARY

In this section, we introduce the mathematical notation used for our proposed algorithm. X is the set of samples and x_j is the j^{th} sample in X. We use $\mathcal{M}$ to denote the set of all must-link pairs and $\mathcal{C}$ to denote the set of all cannot-link pairs. In the application setting, We consider a query by a pair of samples $\langle x_j, x_k \rangle$, And then the answer is $\langle x_j, x_k \rangle \in \mathcal{M}$ or $\langle x_j, x_k \rangle \in \mathcal{C}$. The pairwise constraints satisfy the following properties:

1) If $\langle x_j, x_k \rangle \in \mathcal{M}$, $\langle x_k, x_h \rangle \in \mathcal{M}$, then $\langle x_j, x_h \rangle \in \mathcal{M}$
2) If $\langle x_j, x_k \rangle \in \mathcal{M}$, $\langle x_k, x_h \rangle \in \mathcal{C}$, then $\langle x_j, x_h \rangle \in \mathcal{C}$

B. PROBLEM FORMULATION

In addition to the similarity between samples, the pairwise constraint is an additional clustering principle for the constraint-based clustering algorithm. Fig. 1 provide the operational principle of pairwise constraint. Obviously, it is unnecessary to reconsider the cluster label of samples which are easy to be correctly grouped. If the pairwise constraints consist of these samples, they have less effect on clustering performance. In contrast, the pairwise constraints should work for reconsidering the samples that are most likely to be



(a) Without pairwise constraints **(b)** With pairwise constraints

FIGURE 1. **The operational principle of pairwise constraints in constraint-based clustering. The samples are re-clustered with the help of $\langle h, b\rangle \in \mathcal{M}$, $\langle r, s\rangle \in \mathcal{M}$ and $\langle g, n\rangle \in \mathcal{C}$. They are informative and the clustering performance has improvement. However, $\langle f, k\rangle \in \mathcal{M}$ and $\langle p, x\rangle \in \mathcal{M}$ have less influence on clustering and are non-informative.**

in non-corresponding clusters, which will lead to satisfactory clustering results.

Definition 1: If the pair leads to more desirable sample clustering, it is an informative pairwise constraint (see $\langle h, b\rangle \in \mathcal{M}$, $\langle r, s\rangle \in \mathcal{M}$ and $\langle g, n\rangle \in \mathcal{C}$ in Fig. 1). If the pair cannot improve the clustering performance, it is a non-informative pairwise constraint (see $\langle f, k\rangle \in \mathcal{M}$ and $\langle p, x\rangle \in \mathcal{M}$ in Fig. 1)

C. METHODOLOGY

This section introduces the active informative pairwise constraint formulation algorithm (AIPC) to address the problem of how to efficiently formulate more informative pairwise constraints from limited queries for user application requirements. In our method, the weak sample (see **Definition** 2) and the strong sample (see **Definition** 3) form the informative pair. The queries follow the second minimal symmetric relative entropy priority principle (see **Definition** 4 and **Definition** 5). Fig. 2 provides the overview of the AIPC for constraint-based clustering.

1) PRE-CLUSTERING PHASE

The main purpose of this phase is to obtain the membership degree of each sample and the cluster center of each cluster. In fuzzy c-means clustering (FCM), the membership degree indicates the fuzzy belonging of samples, and the cluster centers represent the pattern of clustering. The objective function of FCM is as follows:

$$J_{FCM}(U, V; X) = \sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^{m} d^2(x_j, \upsilon_i)$$
$$s.t.\ 0 \le \mu_{ij} \le 1, \quad \sum_{i=1}^{c} \mu_{ij} = 1$$
$$1 \le i \le c, \quad 1 \le j \le n \tag{1}$$

where m ($m > 1$) denotes the degree of fuzziness. x_j is the j^{th} sample in samples set X which have n samples. $d(x_j, \upsilon_i)$ can be specified as the Euclidean distance between x_j and v_i. $U = [\mu_{ij}]$ with μ_{ij} being the membership degree of x_j in i^{th} cluster.$V = [\upsilon_1, \upsilon_2, \cdots, \upsilon_c]$ with υ_i being the i^{th} cluster center (c is the number of clusters). By minimizing $J_{FCM}(U, V; X)$, we have the alternative iterative equations in the following

$$\upsilon_i^{(t+1)} = \frac{\sum_{j=1}^{n} \left[\mu_{ij}^{(t)}\right]^m x_j}{\sum_{j=1}^{n} \left[\mu_{ij}^{(t)}\right]^m} \tag{2}$$

and

$$\mu_{ij}^{(t+1)} = \frac{d^2(x_j, \upsilon_i^{(t)})^{\frac{1}{m-1}}}{\sum_{i=1}^{c} d^2(x_j, \upsilon_i^{(t)})^{\frac{1}{m-1}}} \tag{3}$$

The iterations will stop when $\left|J_{FCM}^{(t+1)} - J_{FCM}^{(t)}\right| \le \epsilon_0$ (ϵ_0 is admissible error) or t is equal to T (t and T are the number of iterations and the maximum number of iterations, respectively).

Algorithm 1 *Pre-Clustering* of *AIPC*

Input: X: the set of samples; c: the number of clusters; ϵ_0: the admissible error; m: the degree of fuzziness; T: the maximum number of iterations.

Output: $U = \left[\mu_{ij}^{(t+1)}\right]$: the membership degree; $V = \left[v_1^{(t+1)}, v_2^{(t+1)}, \cdots, v_c^{(t+1)}\right]$: the cluster center.

1: Initial $U = \left[\mu_{ij}^{(0)}\right]$, and iterations number $t = 0$;
2: **repeat**
3: Update cluster center $\upsilon_i^{(t)}$ by (2);
4: Update membership degree $\mu_{ij}^{(t)}$ by (3);
5: Iterations number t++
6: **until** $\left|J_{FCM}^{(t+1)} - J_{FCM}^{(t)}\right| \le \epsilon_0$, or T=t

For the Pre-clustering phase, the FCM provides cluster knowledge that is considered in the Marking phase. The performance of fuzzy clustering will greatly influence the performance of AIPC. This is an important direction of our future research. Our method builds on the membership degree and the cluster center, where the goal is to make use of cluster knowledge. This is different from the above-mentioned selection algorithms which exploit features. In the latter case., the AIPC enters the Marking phase.

2) MARKING PHASE

Definition 2: Suppose that x_j is denoted by samples set X which is grouped into c clusters, and $\{\mu_{1j}, \mu_{2j}, \cdots, \mu_{cj}\}$ is the membership degree of x_j. For $\forall i \in \{1, 2, \cdots, c\}$, there exist $\delta > 0$ (δ is a small enough positive number), such that if

$$\left|\mu_{ij} - \frac{1}{c}\right| < \delta \tag{4}$$

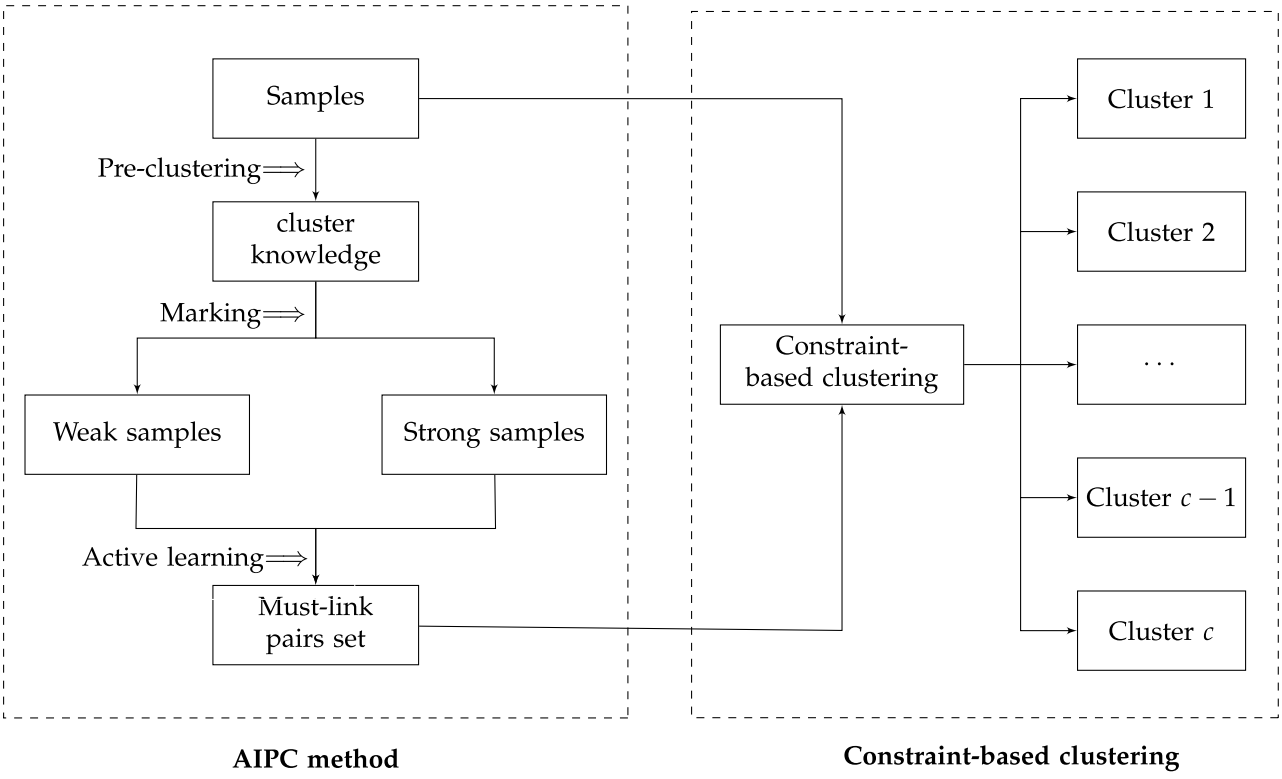


FIGURE 2. The overview of *AIPC* method for constraint-based clustering.

then x_j is a weak sample.

$$E(x_j) = -\sum_{i=1}^{c} \mu_{ij} \ln \mu_{ij} \tag{5}$$

where μ_{ij} is the membership degree of the x_j belonging to the i^{th} cluster. $\mu_{ij} = \frac{1}{c}$ denotes that the uncertainty information is the maximum and the cluster label of x_j is likely to be non-corresponding. Meanwhile, the Shannon entropy is maximum.

$$\max E(x_j) = \ln c \tag{6}$$

Theorem 1: For any $x_j \in X$, there exist $\epsilon > 0$ (ϵ is a small enough positive number), such that

$$E(x_j) > \ln c - \epsilon \tag{7}$$

then x_j is a weak sample.

Proof: For any sample x_j, the membership degree $\{\mu_{1j}, \mu_{2j}, \ldots, \mu_{cj}\}$ is supported in finite dimensional space, and the generalized entropy of x_j is

$$S_f(x_j) = \sum_{i=1}^{c} \mu_{ij} f(\mu_{ij}) \tag{8}$$

where $f : [0, 1] \to [0, \infty)$ is a continuous function with $f(1) = 0$. Let $f(\mu_{ij}) = -\ln \mu_{ij}$, then, the generalized entropy is Shannon entropy.

$$S_f(x_j) = E(x_j) = -\sum_{i=1}^{c} \mu_{ij} \ln \mu_{ij} \tag{9}$$

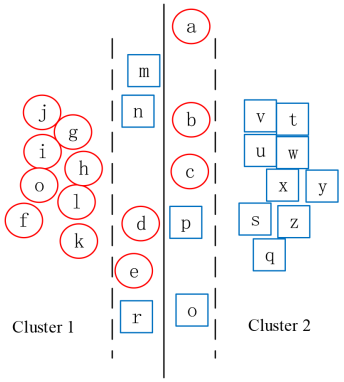


FIGURE 3. Weak sample. The weak samples are between dotted lines.

Obviously, the Shannon entropy is uniform continuity [35].

Thus, for any $\epsilon > 0$, there exist $\delta > 0$ such that for every x_j, x_k with $|\mu_{ij} - \mu_{ik}| < \delta$, we have that

$$|E(x_j) - E(x_k)| < \epsilon \tag{10}$$

Let the membership degree of x_k be $\left\{\frac{1}{c}, \frac{1}{c}, \ldots, \frac{1}{c}\right\}$, $\left|\mu_{ij} - \frac{1}{c}\right| < \delta$, then

$$|E(x_j) - \ln c| < \epsilon \tag{11}$$

Thus, $E(x_j) > \ln c - \epsilon$, x_j is a weak sample (see Fig. 3)

Definition 3: Strong sample is one which contains less uncertainty. Furthermore, it should be closest to the cluster

center of its cluster so that it would be difficult to group into non-corresponding clusters. Each cluster has one strong sample at least.

We apply the membership degree to build up the weak samples set $W = \{x_{w_j} | 1 \le j \le n_w\}$ (n_w is the number of weak samples). The strong sample set $S = \{x_{s_j} | 1 \le j \le n_s\}$ (n_s is the number of strong samples) can be presented by making use of the above cluster centers. Given the weak sample x_{w_j} denoted by W in descending order of entropy, we select one of the strong samples x_{s_j} to formulate the query $\langle x_{w_j}, x_{s_j} \rangle$, where $x_{s_j} \in S$. If x_{w_j} and x_{s_j} are placed in the same cluster, the answer is $\langle x_{w_j}, x_{s_j} \rangle \in \mathcal{M}$.

Although most informative pairs are selected, the number of queries is still larger. We consider the query cost to reach the user's requirement in our method. We define the following processes.

Definition 4: Symmetric relative entropy is developed from relative entropy. Different from relative entropy, it is symmetrical and can more profoundly measure divergence between two samples. Suppose that x_j and x_k are samples in the dataset. Let μ_{ij} be the membership degree of the j^{th} sample belonging to the i^{th} cluster, and μ_{ik} be the membership degree of the k^{th} sample belonging to the i^{th} cluster. The function of symmetric relative entropy is defined as follows:

$$D(x_j, x_k) = \frac{1}{2}\left(\sum_{i=1}^{c} \mu_{ij} \ln \frac{\mu_{ij}}{\mu_{ik}} + \sum_{i=1}^{c} \mu_{ik} \ln \frac{\mu_{ik}}{\mu_{ij}}\right) \quad s.t.\ 1 \le j, \quad k \le n \tag{12}$$

where c is the number of clusters and n is the number of samples in dataset.

Definition 5: Second minimal symmetric relative entropy priority principle is one in which we first select a strong sample whose symmetric relative entropy to weak sample is second minimal.

Following the second minimal symmetric relative entropy priority principle, we first sort the strong samples in ascending order of symmetric relative entropy. Then we formulate the query in the form of: should the $\langle x_{w_j}, x_{s_2} \rangle$ be must-link? If the answer is "Yes", we attain $\langle x_{w_j}, x_{s_2} \rangle \in \mathcal{M}$ and can stop with only one query. If the answer is "No", we get $\langle x_{w_j}, x_{s_2} \rangle \in \mathcal{C}$. In general, the smaller symmetric relative entropy between weak and strong samples, the higher probability of must-link is. Therefore, The strong sample $x_{s_1} \in S$ ($j \ne k, 1 \le k \le n_s$), with minimal symmetric relative entropy to x_{w_j}, formulate $\langle x_{w_j}, x_{s_1} \rangle \in \mathcal{M}$. These processes are summarized in Fig. 4.

For the Marking phase, the query formulates by weak sample and strong sample, leading to informative pairwise constraints. The second minimal symmetric relative entropy priority principle efficiently determines pairwise constraints.

III. EXPERIMENTS

In this section, we will systematically evaluate the performance of AIPC in comparison with four comparative methods. The experiments, conducted in six standard datasets,

Algorithm 2 *Marking* of *AIPC*

Input: X: the set of samples; N_q: the maximum number of queries; $U = [\mu_{ij}]$: the membership degree of each sample; $V = [\upsilon_1, \upsilon_2, \cdots, \upsilon_c]$: the cluster center of each cluster.

Output: The informative must-link pairwise constraints

1: the number of queries $n_q = 0$;
2: Calculate the entropy of membership degree and then get the weak samples;
3: Sort weak samples $W = \{x_{w_1}, x_{w_1}, \ldots, x_{w_{n_w}}\}$ in descending order of entropy;
4: Select strong samples which are closest to their cluster centers;
5: **repeat**
6: Select weak sample $x_{w_j} \in W$;
7: Sort strong samples $S = \{x_{s_1}, x_{s_1}, \ldots, x_{s_{n_s}}\}$ in ascending order of symmetric relative entropy between strong and weak samples
8: seek answer to the query $\langle x_{w_j}, x_{s_2} \rangle$
9: If the answer is must-link, $\langle x_{w_j}, x_{s_2} \rangle \in \mathcal{M}$ is returned; If the answer of is cannot-link, $\langle x_{w_j}, x_{s_1} \rangle$ be specified as must-link(see Fig. 4);
10: the number of queries n_q++
11: **until** $N_q \le n_q$

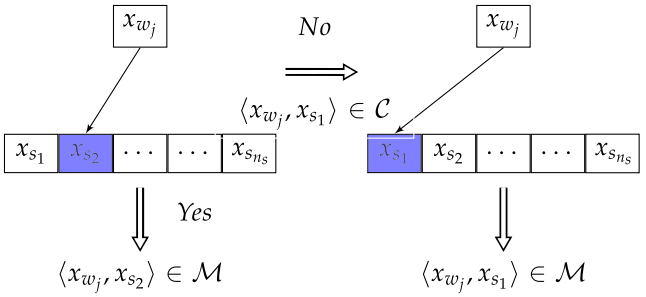


FIGURE 4. The overview of second minimal symmetric relative entropy priority principle. The x_{s_1} denotes the strong sample whose symmetric relative entropy between the j^{th} weak sample x_{w_j} is minimal.

introduce PCKM as the underlying constraint-based clustering algorithm. There are two performance metrics used in clustering. The experimental results demonstrate that our AIPC method has substantial improvements.

A. COMPARATIVE METHODS

To evaluate what AIPC brings in term of improvements of the performance, we compare AIPC with a set of comparative methods, including Random algorithm, FFQS algorithm [8], Min-Max algorithm [9], and Cai algorithm [13]. The Random is the baseline method. The FFQS algorithm is the classic method which is in query view. The Min-Max, an important variant of FFQS, is the popular comparative method of the pairwise constraint formulation algorithm. The Cai algorithm is an important research study in sample view. In the following, we will briefly explain these comparative methods:

1) The Random algorithm is a method which randomly selects the samples to form the pairwise constraints. The cost of the query is high. In the meanwhile, there may be non-informative pairs that lead to poor clustering performance.
2) The FFQS algorithm [8] is made up of two phases: Explore and Consolidate phase. The Explore phase selects at least one neighborhood sample from each cluster by using the farthest-first strategy. These samples may have corresponding cluster labels. The Consolidate phase randomly selects the sample not included in neighborhoods set and formulates queries against neighborhoods in increasing order of distance until a must-link pair is obtained.
3) The Min-Max algorithm [9], which builds on the Explore phase of FFQS [8], improves the Consolidate phase. They select samples in which the largest similarity to the neighborhoods sample is the smallest. Then they formulate the query again with the neighborhood sample that it is closest to. The query is answered in descending order of similarity until a must-link query is obtained.
4) The Cai method [13], which is inspired by the k-nearest neighbors (KNN) algorithm, gets an informative sample set to study the formulation algorithm. Then the Explore and Consolidate phase work on these informative samples. In Explore phase, the most informative sample is considered as the first sample. The rest of the procedure is the same as Min-Max [9].

B. DATASETS

In our experiments, we use six standard UCI datasets which are popular for evaluating the performance of pairwise constraint formulation [8], [9], [13]. They include Iris, Wine, Letters recognition (IJLTY), Pen-based recognition of handwritten digits (3,8,9) (briefly Digits-389) and Ecoli. For the Ecoli dataset, we remove the smallest 3 clusters, which is used in common. The characteristics of them are in following Table 2.

TABLE 2. The characteristics of datasets.

Datasets	# of Instances	# of Attributes	# of Clusters
Iris	150	4	3
Wine	178	13	3
Letters-IJLTY	3845	16	5
Breast	683	9	2
Digits-389	3165	16	3
Ecoli	327	7	5

C. UNDERLYING CONSTRAINT-BASED CLUSTERING

The clustering result of underlying constraint-based semi-supervised clustering algorithm relates to the performance of the pairwise constraint formulation algorithm. Our proposed method AIPC and other comparative methods assume the validity of constraint-based semi-supervised clustering. For this purpose, we use PCKM as the implementation of the above-mentioned pairwise constraint formulation algorithm. Although some researchers have proposed a lot of constraint-based semi-supervised clustering derived from PCKM [2], [3], the PCKM is one which is effective and easy to implement. The PCKM, a K-means variant, incorporates prior information in the form of pairwise constraint which is an additional clustering principle. Different from the classic unsupervised clustering such as k-means, it minimizes not only the total distance between samples and the cluster centers but also the cost of violating the pairwise constraints. The objective function is in following:

$$J_{PCKM}(L, V; X) = \sum_{i=1}^{c}\sum_{j=1}^{n} d^2(x_j, \upsilon_i) + \sum_{\langle x_j, x_k\rangle \in \mathcal{M}} \omega_{jk}\, \mathbb{1}\left[l'_j \neq l'_k\right] + \sum_{\langle x_j, x_k\rangle \in \mathcal{C}} \overline{\omega}_{jk}\, \mathbb{1}\left[l'_j \neq l'_k\right] \quad (13)$$

where x_j is the j^{th} sample of dataset X. υ_i is the i^{th} cluster center of V. $L' = \left[l'_j\right]$ with l'_j denotes the cluster label of x_j. $\mathbb{1}$ is the indicator function, $\mathbb{1}[ture] = 1$ and $\mathbb{1}[false] = 0$. $W = \left[\omega_{jk}\right]$ and $\overline{W} = \left[\overline{\omega}_{jk}\right]$ is penalty cost for violating the $\langle x_j, x_k\rangle \in \mathcal{M}$ and $\langle x_j, x_k\rangle \in \mathcal{C}$ respectively.

In the experiment, we set the maximum times of iterations of PCKM to 100 and use the default values for other parameters. We formulate up to 100 pair queries, starting from no constraint at all. The time complexity of PCKM is $O(n)$. The time complexity of our method *AIPC* and other comparative methods implementing in PCKM are shown in Table 3. It is obvious that the complexity of our implementation is acceptable. All of implementation run on unified experimental setting platform. The characterizations of it is shown in Table 4.

TABLE 3. The summary of time complexity.

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

D. PERFORMANCE METRIC

For comparative analysis, there are two clustering performance metric applied in the experiments. First, the Rand index (RI) is a popular clustering evaluation metric [36]–[38]. The value of RI is between 0 and 1. A value close to 1 indicates that the performance is desirable. And the value equal to 0 shows that the data clusters are completely different.

TABLE 4. The characterizations of platform.

Item	Characterizations
Program	MATLAB 2018a
Memory	8.00G
CPU	Intel(R) Core(TM) i7-7700
Operate system	Windows 10 professional

TABLE 5. The values of ϵ for weak sample.

Datasets	Iris	Wine	Letters-IJLTY
ϵ	9.0×10^{-1}	8.0×10^{-1}	2.1×10^{-6}
Datasets	Breast	Digits-389	Ecoli
ϵ	1.4×10^{-1}	9.0×10^{-3}	1.1×10^{-1}

Suppose that the dataset $X = \{x_1, x_2, \cdots, x_n\}$ with its actual cluster label set $L = \{l_1, l_2, \cdots, l_n\}$. After clustering, we get predicted cluster label set $L' = \{l'_1, l'_2, \cdots, l'_n\}$. We have definition in following.

$$TP = \left\{(x_j, x_k)|, l_j = l_k, l'_j = l'_k, j \neq k\right\} \tag{14}$$

$$FP = \left\{(x_j, x_k)|, l_j = l_k, l'_j \neq l'_k, j \neq k\right\} \tag{15}$$

$$TN = \left\{(x_j, x_k)|, l_j \neq l_k, l'_j \neq l'_k, j \neq k\right\} \tag{16}$$

$$FN = \left\{(x_j, x_k)|, l_j \neq l_k, l'_j = l'_k, j \neq k\right\} \tag{17}$$

The objective function of RI is

$$RI = \frac{|TP| + |TN|}{|TP| + |FP| + |FN| + |TN|} \tag{18}$$

Second, the normalized mutual information (NMI) [39], [40] is used to measure the mutual information between predicted cluster label set and actual cluster label set. It is normalized to a zero-to-one range. The equation is as follows.

$$NMI = 2\frac{I(L'; L)}{H(L') + H(L)} \tag{19}$$

where $I(L'; L)$ is the mutual information between L' and L, $H(L')$ denotes the entropy of L'.

E. IMPLEMENTATION DETAILS

There are two experiments being implemented in this paper: 1) introducing pairwise constraint formulation algorithm (our AIPC and other comparative algorithms) to PCKM; 2) performing sensitivity analysis for the parametric ϵ of our AIPC.

In the first experiment, we set the maximum times of iterations T of FCM to 100 and use the admissible error $\epsilon_0 = 10^{-5}$ in the Pre-clustering phase of AIPC. The degree of fuzziness m takes 2, which is more popular in FCM applications. According to **Theorem 1**, the ϵ would provide the equivalent of δ. And it makes the AIPC more effective and operational. There are three phases about how to define ϵ: 1) sort samples in descending order of entropy; 2) confirm the number of queries and clusters as n_q and c respectively; 3) the value of ϵ is $\ln c - E(x_{n_q})$ where $E(x_{n_q})$ denotes the entropy of the n_qth sample. The ϵ of this experiment are shown in Table 5.

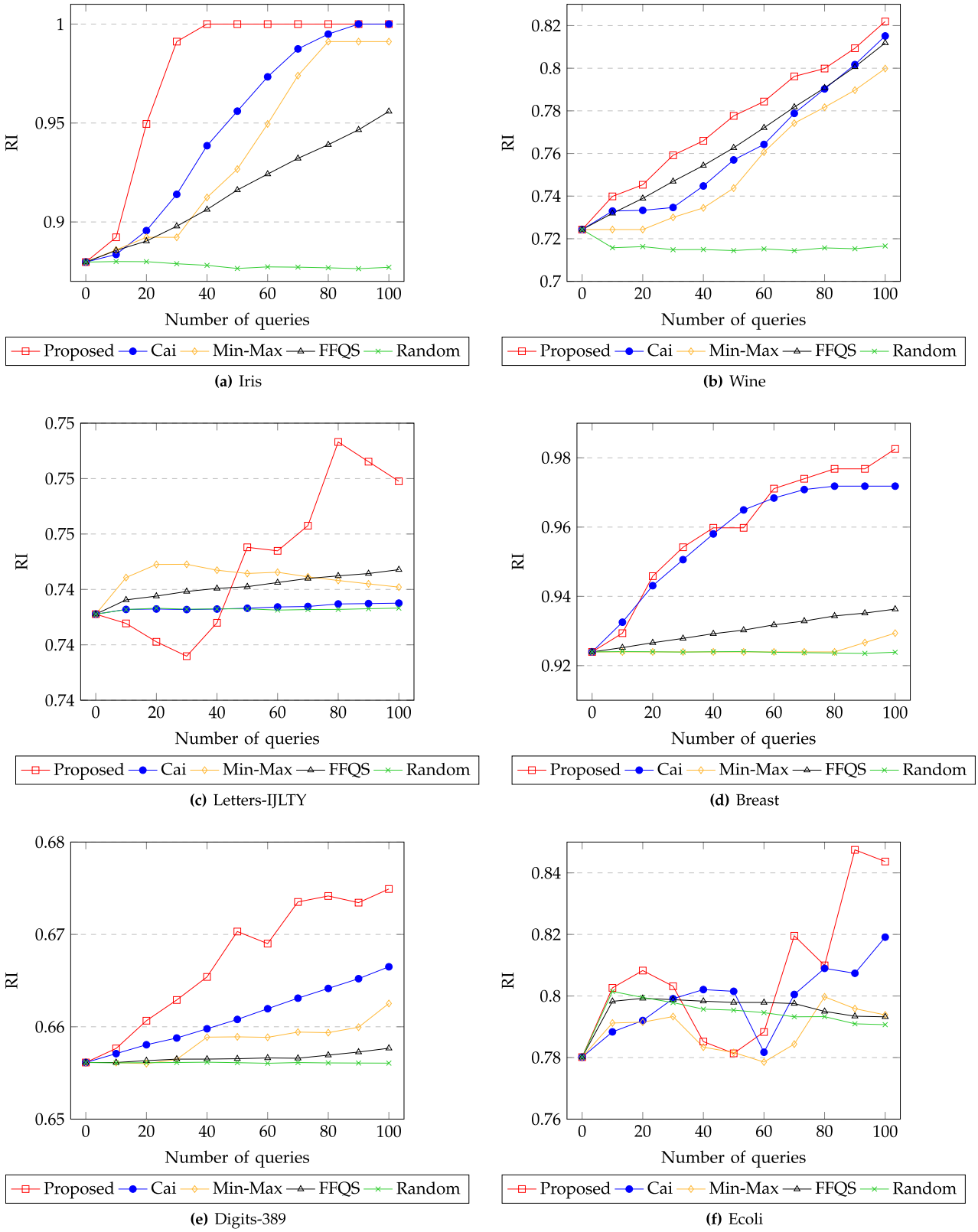
In the second experiment, the number of queries takes from $n_q = 0$ to $n_q = 100$, we evaluate the influence of ϵ with its six different values.

F. EXPERIMENTAL RESULTS AND THEIR ANALYSIS

In this section, we present the clustering result of PCKM which are in conjunction with Random, FFQS, Min-Max, Cai and our proposed method on six different datasets. The result is composed of RI and NMI value which is the mean of 50 independent runs. In general, the AIPC outperforms other comparative methods.

The RI and NMI index of clustering result are shown in Fig. 5 and Fig. 6 respectively. The x-axis indicates the number of queries and the y-axis presents the RI or NMI by running PCKM with 4 comparative methods and our proposed. The more powerful the clustering performance of PCKM, the more desirable result of the pairwise constraint formulation algorithm we attain. Note that the pairwise constraint provided by Random algorithm lead to terrible clustering performance. In comparison, the other comparative method and AIPC are generally able to improve the performance with the increasing of the number of queries. However, the AIPC are more powerful than other formulation algorithm in most case.

In Iris, one cluster is linearly separable from the other two clusters. The RI and NMI value are close to one consistently as we increase the number of queries. The AIPC, Cai and Min-Max converge before using up 100 queries. But AIPC requires fewer queries to obtain the same result. In Wine, the AIPC method obtains better results than FFQS and Cai method through lager query sizes. It is noteworthy that AIPC degrades the clustering result in Letters-IJLTY dataset when the queries are in the early stage. This demonstrates that some formulated pairwise constraints have inaccurate labels. When all labeled pairs are introduced into PCKM, the positive what the correct labeling pairs bring to can hardly offset the negative of the wrong labeling pairs. But as we increase the number of queries, the advantage of the role of correct labeling pair began to appear gradually and the performance of AIPC becomes better and better. We also note that the performance strength of our proposed is obvious in some dataset, namely Breast, Digits-389. For Ecoli, there are 5 clusters. The RI value are low when the queries work in the early stage. While the number of queries increases to a certain extent, the clustering performance has rapid development. In other words, the performance of AIPC may be not significant with comparison of other methods when the query size is small. If the queries continue, the AIPC shows better performance.

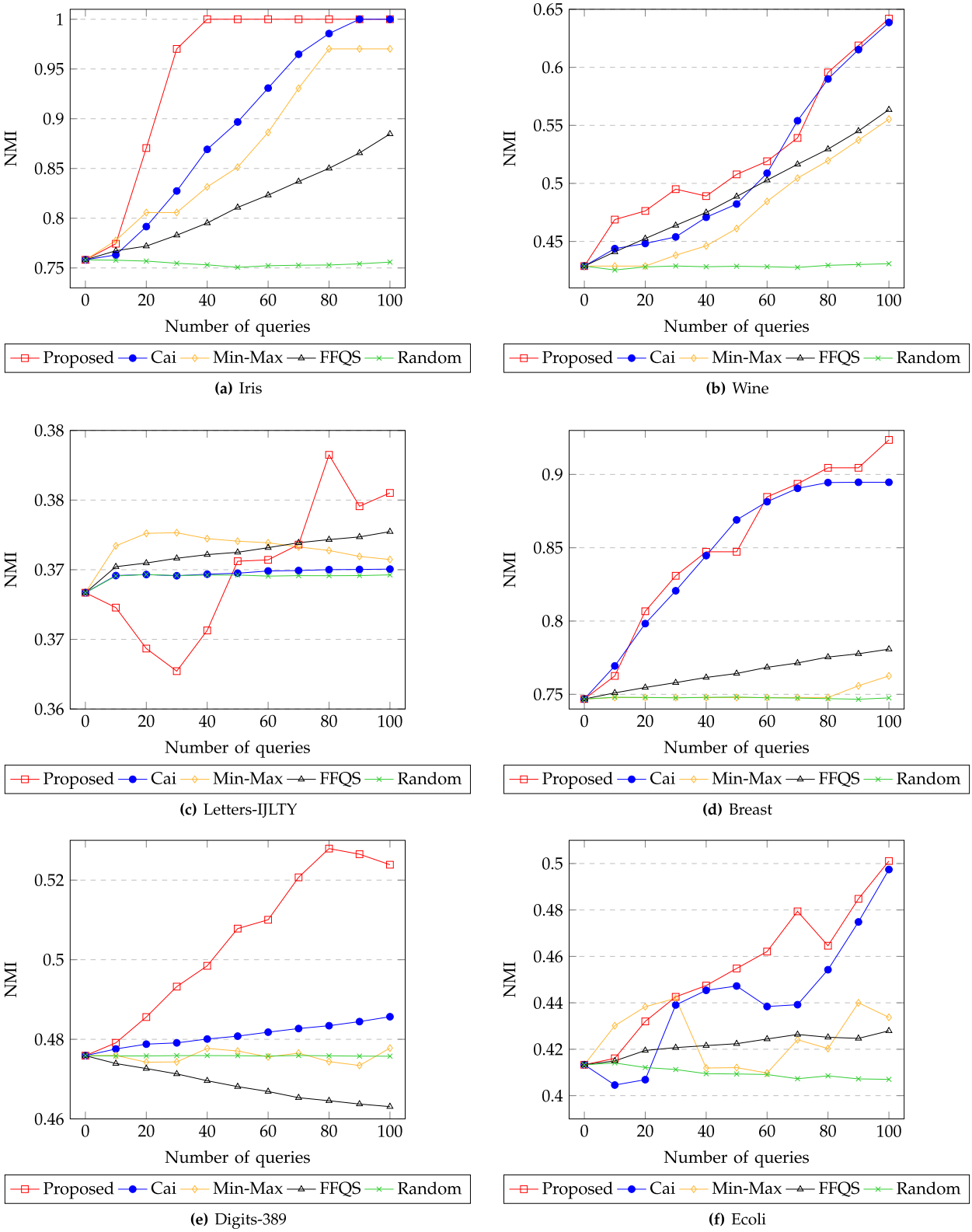


(a) Iris **(b)** Wine **(c)** Letters-IJLTY **(d)** Breast **(e)** Digits-389 **(f)** Ecoli

FIGURE 5. Evaluation results of clustering performance by RI index.

G. SENSITIVITY ANALYSIS

In order to understand what are the AIPC algorithm parametric ϵ contributing to our proposed, we provide the sensitivity analysis experiments which are conducted in three standard UCI datasets: Breast, Digits-389, Letters-IJLTY. They have different numbers of classes and contain the largest number of

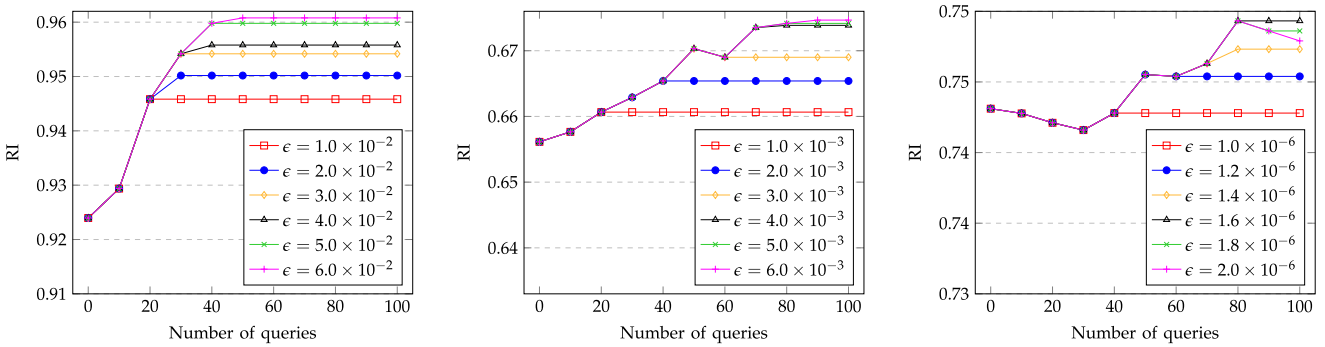


(a) Iris **(b)** Wine **(c)** Letters-IJLTY **(d)** Breast **(e)** Digits-389 **(f)** Ecoli

FIGURE 6. **Evaluation results of clustering performance by NMI index.**

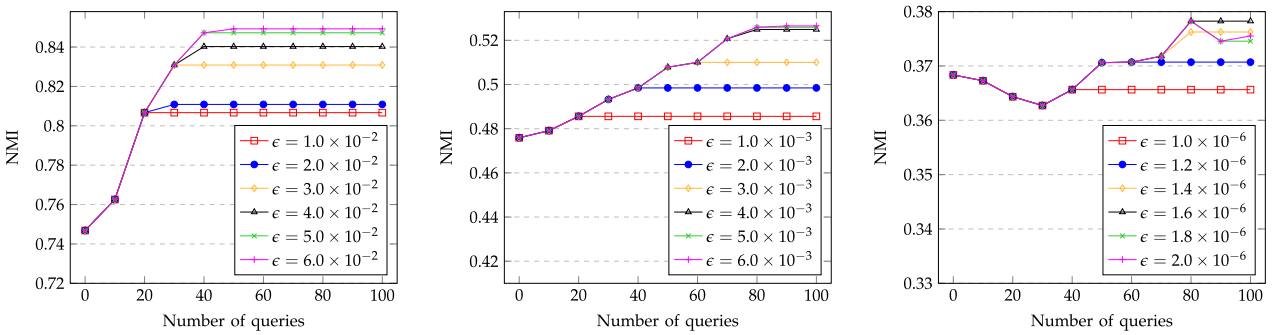
samples. Specifically, the experiment selects different values of ϵ (see Table 6) to study the effect of it on AIPC. The experimental result of sensitivity analysis is shown in Fig. 7 and Fig. 8

From the horizontal view, the performance of our method becomes more and more powerful when we had an increasing number of queries. If the RI/NMI stop rising, it indicates that the number of weak samples is less than the number of



(a) Breast **(b)** Digits-389 **(c)** Letters-IJLTY

FIGURE 7. The PCKM clustering result of RI with introducing AIPC which considers 6 different values of ϵ.



(a) Breast **(b)** Digits-389 **(c)** Letters-IJLTY

FIGURE 8. The PCKM clustering result of NMI with introducing AIPC which considers 6 different values of ϵ.

queries. From the vertical view, our method gets more and more desirable result with the increasing of ϵ. In the meanwhile, the improvement degree becomes less and less. Integrally, it is necessary for selecting the appropriate value of ϵ to consider the cost of computation and classifier performance.

IV. CONCLUSION AND FUTURE WORK

In the Pre-clustering phase, our goal is to obtain the membership degree and cluster center. The clustering performance of fuzzy clustering has a vital impact on the performance of AIPC. The more powerful the cluster knowledge, the more desirable the performance we obtain. The FCM is one of the most widely used fuzzy clustering algorithms. A great number of problems from different application scenes have been effectively solved by introducing FCM [41]–[44]. In the future, we will research developing fuzzy clustering to improve AIPC

In the Marking phase, our goal is to form informative pairs and make queries more efficient. The Shannon entropy is one measure to describe the uncertainty of samples. The most informative pair should be composed of two weak samples. However, it is difficult for experts to answer. In the future, this will be a research direction.

REFERENCES

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

[2] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proc. Int. Conf. Mach. Learn.*, 2004, p. 11.

[3] N. Grira, M. Crucianu, and N. Boujemaa, "Active semi-supervised fuzzy clustering," *Pattern Recognit.*, vol. 41, no. 5, pp. 1834–1844, May 2008.

[4] Z. Yu, Z. Kuang, J. Liu, H. Chen, J. Zhang, J. You, H. S. Wong, and G. Han, "Adaptive ensembling of semi-supervised clustering solutions," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 8, pp. 1577–1590, Aug. 2017.

[5] G. R. da Silva and M. K. Albertini, "Using multiple clustering algorithms to generate constraint rules and create consensus clusters," in *Proc. Brazilian Conf. Intell. Syst. (BRACIS)*, Uberlandia, Brazil, Oct. 2017, pp. 312–317.

[6] Z. Yu, P. Luo, J. Liu, J. You, H.-S. Wong, 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.

[7] K. Wagstaff and C. Cardie, "Clustering with instance-level constraints," in *Proc. Int. Conf. Mach. Learn.*, 2000, pp. 1103–1110.

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

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

[10] Q. Xu, M. Desjardins, and K. L. Wagstaff, "Active constrained clustering by examining spectral eigenvectors," in *Proc. Int. Conf. Discovery Sci.*, 2005, pp. 294–307.

[11] R. Huang and W. Lam, "Semi-supervised document clustering via active learning with pairwise constraints," in *Proc. IEEE Int. Conf. Data Mining*, Oct. 2007, pp. 517–522.

[12] 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.

[13] L. Cai, T. Yu, T. He, L. Chen, and M. Lin, "Active learning method for constraint-based clustering algorithms," *Web-Age Information Management*. Cham, Switzerland: Springer, 2016.

[14] 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, 2015.
[15] B. Du, Z. Wang, L. Zhang, L. Zhang, and D. Tao, "Robust and discriminative labeling for multi-label active learning based on maximum correntropy criterion," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1694–1707, Apr. 2017.
[16] G. Wang, J.-N. Hwang, C. Rose, and F. Wallace, "Uncertainty sampling based active learning with diversity constraint by sparse selection," in *Proc. IEEE 19th Int. Workshop Multimedia Signal Process. (MMSP)*, Oct. 2017, pp. 1–6.
[17] G. Wang, J.-N. Hwang, C. Rose, and F. Wallace, "Uncertainty-based active learning via sparse modeling for image classification," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 316–329, Jan. 2019.
[18] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, Nov. 2002.
[19] 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.
[20] S. Zhang, G. Deng, and F. Wang, "Active learning strategy for online prediction of particle size distribution in cobalt oxalate synthesis process," *IEEE Access*, vol. 7, pp. 40810–40821, 2019.
[21] J. Chen, D. Zhou, Z. Guo, J. Lin, C. Lyu, and C. Lu, "An active learning method based on uncertainty and complexity for gearbox fault diagnosis," *IEEE Access*, vol. 7, pp. 9022–9031, 2019.
[22] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Proc. 11th Int. Conf. Mach. Learn.*, 1994, pp. 148–156.
[23] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2372–2379.
[24] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proc. 5th Workshop Comput. Learn. Theory*, vol. 284, 1992, pp. 287–294.
[25] R. Burbidge, J. J. Rowland, and R. D. King, "Active learning for regression based on query by committee," in *Proc. 8th Int. Conf. Intell. Data Eng. Automated Learn. (IDEAL)*, Berlin, Germany: Springer-Verlag, 2007, pp. 209–218.
[26] C. Tekin and M. van der Schaar, "Active learning in context-driven stream mining with an application to image mining," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3666–3679, Nov. 2015.
[27] B. Krawczyk and M. Woźniak, "Online query by committee for active learning from drifting data streams," in *Proc. Int. Joint Conf. Neural Netw.*, 2017, pp. 2120–2127.
[28] N. Roy and A. Mccallum, "Toward optimal active learning through sampling estimation of error reduction," in *Proc. 11th Int. Conf. Mach. Learn.*, 2001, pp. 441–448.
[29] 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. Int. Conf. Mach. Learn.*, 2006, pp. 417–424.
[30] Y. Guo and R. Greiner, "Optimistic active learning using mutual information," in *Proc. Int. Joint Conf. Artif. Intell.*, 2007, pp. 823–829.
[31] X. Zhu, P. Zhang, X. Lin, and Y. Shi,, "Active learning from stream data using optimal weight classifier ensemble," *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 40, no. 6, pp. 1607–1621, Dec. 2010.
[32] S. J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 892–900.
[33] H.-M. Chu and H.-T. Lin, "Can active learning experience be transferred?" in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Barcelona, Spain, Dec. 2016, pp. 841–846.
[34] Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway, and J. Liang, "Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7340–7351.
[35] X. Cao and S. Luo, "On the stability of generalized entropies," *J. Phys. A, Math. Theor.*, vol. 42, no. 7, 2009, Art. no. 075205.
[36] G. Gan and K. P. Ng, *K-Means Clustering With Outlier Removal*. Amsterdam, The Netherlands: Elsevier, 2017.
[37] X. Huang, X. Yang, J. Zhao, L. Xiong, and Y. Ye, "A new weighting k-means type clustering framework with an l^2-norm regularization," *Knowl.-Based Syst.*, vol. 151, pp. 165–179, Jul. 2018.
[38] J. Lu and Q. Zhu, "An effective algorithm based on density clustering framework," *IEEE Access*, vol. 5, pp. 4991–5000, 2017.
[39] L. I. Kuncheva and S. T. Hadjitodorov, "Using diversity in cluster ensembles," in *Proc. IEEE Int. Conf. Syst.*, Oct. 2004, pp. 1214–1219.
[40] 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.
[41] L. Liu, C.-F. Li, Y.-M. Lei, J.-J. Zhao, J.-Y. Yin, and X.-K. Sun, "A new fuzzy clustering method with neighborhood distance constraint for volcanic ash cloud," *IEEE Access*, vol. 4, pp. 7005–7013, 2016.
[42] A. Arshad, S. Riaz, L. Jiao, and A. Murthy, "Semi-supervised deep fuzzy c-mean clustering for software fault prediction," *IEEE Access*, vol. 6, pp. 25675–25685, 2018.
[43] K. Wisaeng and W. Sa-Ngiamvibool, "Improved fuzzy c-means clustering in the process of exudates detection using mathematical morphology," *Soft Comput.*, vol. 22, no. 8, pp. 2753–2764, 2018.
[44] A. Feizollah, N. B. Anuar, and R. Salleh, "Evaluation of network traffic analysis using fuzzy c-means clustering algorithm in mobile malware detection," *Adv. Sci. Lett.*, vol. 24, no. 2, pp. 929–932, 2018.

GUOXIANG ZHONG received the B.S. degree from the Guangdong University of Technology, in 2017, where he is currently pursuing the master's degree with the School of Apply Mathematics. His main research interests include semi-supervised learning and computational intelligence.

XIUQIN DENG received the B.S. degree in mathematics from the Minzu University of China, Beijing, in 1987, and the M.S. degree in computer application from the South China University of Technology, Guangzhou, China, in 2005. She is currently a Professor with the School of Apply Mathematics, Guangdong University of Technology. Her research interests include data mining, machine learning, and their applications.

SHENGBING XU received the B.S. degree in computational mathematics and applied software and the M.Sc. degree in applied mathematics from Xiangtan University, in 1997 and 2001, respectively. He is currently pursuing the Ph.D. degree with the Guangdong University of Technology. His research interests include fuzzy set, mathematical modeling, machine learning, and their applications. He is also a Peer Reviewer of IEEE Access.