IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 3, 2020, accepted November 20, 2020, date of publication November 26, 2020,
date of current version December 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3040745

An Improved K-Means Algorithm Based on Fuzzy Metrics

XINYU GENG, YUKUN MU, SENLIN MAO, JINCHI YE, AND LIPING ZHU
School of Computer Science, Southwest Petroleum University, Chengdu 610500, China
Corresponding author: Xinyu Geng (gengxy123@126.com)

ABSTRACT The traditional K-means algorithm has been widely used in cluster analysis. However, the algorithm only involves the distance factor as the only constraint, so there is a problem of sensitivity to special data points. To address this problem, in the process of K-means clustering, ambiguity is introduced as a new constraint condition. Hence, a new membership Equation is proposed on this basis, and a method for solving the initial cluster center points is given, so as to reduce risks caused by random selection of initial points. Besides, an optimized clustering algorithm with Gaussian distribution is derived with the utilization of fuzzy entropy as the cost function constraint. Compared with the traditional clustering method, the new Equation's membership degree can reflect the relationship between a certain point and the set in a clearer way, and solve the problem of the traditional K-means algorithm that it is prone to be trapped in local convergence and easily influenced by noise. Experimental verification proves that the new method has fewer iterations and the clustering accuracy is better than other methods, thus having a better clustering effect.

INDEX TERMS K-means, fuzzy entropy, cluster center, membership degree, fuzzy clustering.

I. INTRODUCTION

The clustering process is the most effective classification method for people to summarize complex external information [1]. Though classification can see a mature development now, there are still challenges for the clustering algorithm regarding how to eventually realize cognition, learning and classification under unsupervised conditions by extracting data features [2]. No model can be used universally and achieve better results, since it is not a priori [3]. Data imply enormous scientific and commercial values [4], especially in the explosive growth of data production in recent years. In 2016, the global data volume reached 10ZB and maintained an annual growth rate of more than 40 Scattered raw data, processed with data mining technology, can deliver valuable results, such as the planning of humanities and the construction of biological sciences in the reference [6]–[8]. This type of research is of great significance for both social development and human self-cognition and learning cognition. It can be clearly seen that clustering research on various types of data has attracted academic attention for a long time [9].

In a study on clustering problems, for a given data set, we should first make sure whether there is a clustering structure. If so, the algorithm structure should be determined. Once conformed, three aspects should be involved to figure out whether the clustering result is reasonable or not [10]. For different types of data, there are different processing algorithms to get a good clustering effect. K-means is one of the most commonly used clustering algorithms. Featuring a simple principle, the K-means algorithm is easy to implement, and can classify low-dimensional large data sets in an efficient way. However, the K-means algorithm also has shortcomings such as the vulnerability to special points, the risk of local optimal solutions, the only constraint of distance, the sensitivity to initial point selection, and the exclusive suitability for clustering numerical data and data sets with convex clusters [11]. Therefore, based on the classic K-means algorithm, many new and improved algorithms have been proposed, such as the K-modes algorithm that can cluster discrete data [12], K-means-CP algorithm based on the consistency of k nearest neighbors [13], Canopy-based K-means (DCK-means) algorithm with Canopy Method integrated through which the initial point is found by considering sample density [14] etc. The K-means algorithm and its derived algorithms have been successfully applied in recommendation systems, image processing, data mining, video recognition, and other fields [15]. At present, the classic K-means algorithm is usually optimized by combining other algorithms to realize the selection of clustering centers and the determination of the distance function In [16], it is proposed to use the result of Singular Value Decomposition (SVD) decomposition as the initial point of clustering to

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

obtain better clustering effects. In reference [17] it is presented to utilize neighboring ideas, by taking the overall distribution of data samples as the basis for division, so as to improve clustering effects. In reference [18], with the combination of Adaptive radius immune algorithm(ARIA) and K-means algorithms, the immune algorithm for adaptive radius is applicated to preprocess the data so as to generate mirror data that represen the distribution and density of the original data, improving the noise resistance and stability of clustering. In reference [19], a K-means clustering algorithm is designed with a random shift of the center of gravity to process single-color images specifically for the isolated point problem. In addition, new directions have emerged, which combine swarm intelligence and bionic algorithms, such as Particle swarm optimization K-means(PSO-Kmeans) based on particle swarms [20], artificial bee colony K-means algorithm (ABC-Kmeans) based on bee swarms [21], and Gray Wolf K-means (GWO-Kmeans) [22] based on gray wolf optimization. In the above optimization examples, most are still optimization methods for the division of a certain aspect such as distance or edge data. Among them, only reference [18] presented an application for the characteristics of data distribution, which is an effective use of data information except for the distance between points.

After the fuzzy theory appeared, Dunn *et al.* put forward the fuzzy c-means clustering algorithm (FCM) [23]. According to its judgement criteria, clustering is divided into hard clustering and soft clustering. Reference [24] raised a Fuzzy C-Means algorithm with a Divergence-based Kernel FCM algorithm (FCMDK), which can handle data whose boundaries between clusters were non-linear. Chowdhary, C. L *et al.*proposesd a novel possibilistic exponential fuzzy c-means (PEFCM) clustering algorithm for segmenting medical images better and earlier [25]. In the process of hard clustering, objects to be clustered are strictly classified by their exclusive nature, but the object of a practical problem is complex, and there may exist objective problems with attributes of multiple categories. Therefore, it is necessary to come up with a soft division method on such fundamental issues. For example, FCM clustering has been widely used in fields like pattern classification and image processing [26]. Because it adds ambiguity to the membership requirement of each pixel, this type of algorithm is significantly better than traditional K-means clustering in image processing results [27]. Although the degree of membership can feedback the correlation beyond the distance, it is still gained by the distance relationship between a single point and the class. Therefore, the algorithm is also very susceptible to noise. In reference [28] hybridizes intuitionistic fuzzy set and rough set in combination with statistical feature extraction techniques,which is higher than the accuracy achieved by hybridizing fuzzy rough set model. In reference [29], a method is advanced to relax the restrictions of membership and improve the robustness. In reference [30], the weight of membership degree is involved for the consideration of influences of each dimension attribute on clustering. In reference [31], a heuristic method is offered based on the Silhouette criterion to find the number of clusters. Reference [32] designates the digital execution of a model, based on an intuitionistic fuzzy histogram hyperbolization and possibilitic fuzzy c-mean clustering algorithm for early breast cancer detection. In reference [33], a clustering method featuring the combination of fuzzy c-means algorithm and entropy- based algorithm is advised to achieve both distinct and compact effects. Reference [34]proposed a novel intuitionistic possibilistic fuzzy c-mean algorithm. Possibilistic fuzzy c-mean and intuitionistic fuzzy c-mean are hybridized to overcome the problems of fuzzy c-mean. In reference [35], a novel fuzzy-entropy based clustering measure (FECM) is presented, in which the average symmetric fuzzy cross entropy of membership subset pairs is integrated with the average fuzzy entropy of clusters. The above methods improve clustering effects with the fuzzy entropy utilized to enhance the effective use of information such as overall distribution characteristics. Inspired by the theory of fuzzy mathematics and reference [17] and [18], in the clustering process, this paper proposes a fuzzy mean clustering algorithm, namely fuzzy metrics K-means (FMK), with fuzzy entropy as a constraint in the distance condition. The algorithm first introduces artificial setting of the initial cluster center to reduce the influence of noise, integrates the overall distribution structure into that of the membership function, and then compares the overall ambiguity of the cluster after introducing a certain point. The last step is the convergence completed through iteration, finally realizing the clustering of the FMK-means algorithm.

II. RELATED INFORMATION

A. K-MEANS ALGORITHM

As one of the most well-known clustering algorithms, The K-means algorithm can actively select the number of categories and bases its calculation of closeness on the Euclidean distance between the points.

In the algorithm, k clusters $\boldsymbol{C} = \{C_1, C_2, \cdots, C_k\}$are randomly selected as partitions and n datasets of samples $\boldsymbol{D} = \{x_1, x_2, \cdots, x_n\}, n \geqslant k$are divided into the nearest cluster. Then Recalculate the center point of the cluster. Stop the iteration until the convergence condition is reached. Generally, the convergence function is defined as follow:

$$SSE(\boldsymbol{C}) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \| x_i - c_k \|^2 \tag{1}$$

in, SSE is the least square error of the cluster division corresponding to the algorithm sample clustering. c_k is the center point of the cluster C_k. In the reference [15], the mathematical meaning of the center point is verified and deduced. Here comes the conclusion: the best center of a cluster is the mean value of each point in it. The calculation method is as follow:

$$c_k = \frac{\sum_{x_i \in C_k} x_i}{|C_k|} \tag{2}$$

In summary, the goal of the K-means algorithm is the minimum clustering result in (1). This optimization problem

is an NP-hard problem [15]. Equation (1) also reflects the closeness of the samples in the cluster around the center. The smaller the SSE value is, the higher the degree of clustering and the higher the similarity within the cluster will be. The algorithm is characterised by its simple method in generating the center point, fast speed and good scalability.

B. MEMBERSHIP AND FUZZY METRICS

The degree of membership is the basis of fuzzy set operations, and the membership function is the key to describing the fuzziness [36]. Unlike the rule of the classic set, one element in a fuzzy set can belong to multiple sets. Besides, the sum of the membership degrees of the element to different clusters is always one. For example, in the FCM algorithm, the membership algorithm is as follow:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{d_{ij}}{d_{kj}})^{\frac{2}{m-1}}} \quad s.t. \quad \sum_{j=1}^{c} u_{ij} = 1 \tag{3}$$

The relationship between a point and a set is described by the degree of membership, while the overall fuzzy degree of a certain set requires a new fuzzy measure. The fuzzy entropy defined based on the order relationship reflects the fuzzy degree of a fuzzy partition [37]. Let $F(X)$ denote all fuzzy sets on X. For any $A, B \in F(X)$, we call $A \leqslant {}^{F}_{M} B$ if and only if $\bar{A} \leqslant \bar{B}$. There are the following theorems on the ambiguity $d(A)$ of fuzzy set A [34,35]

Theorem 1: Let A be the fuzzy set in the universe $\boldsymbol{X}$, if the mapping d: $F(U) \rightarrow [0, 1]$ then: (1)

1) $d(A) = 0$ if and only if $A \in \rho(X)$;
2) $d(A) = 1$ if and only if $\mu A(x) \equiv 0.5$;
3) If any $x_i \in U$, $\mu A(x_i) \geqslant \mu B(x_i) \geqslant 0.5$, or $\mu A(x_i) \leqslant \mu B(x_i) \leqslant 0.5$, then $d(A) \leqslant d(B)$;
4) $d(A) = d(\sim A)$

Theorem 2: IF $U_1 = [A_1, A_2, \cdots, A_c]$, $U_2 = [B_1, B_2, \cdots, B_c]$ are two fuzzy c divisions on $\boldsymbol{X}$ and satisfy $A_j \leqslant {}^{F}_{M} B_j, j = 1, 2, \cdots, c$,then call U_1 a distinct modification of U_2 and record it as $U_1 \prec {}^{F}_{M} U_2$. Hence, for all fuzzy partitions $FP(X)$ on $\boldsymbol{X}$, there must exist $[\frac{1}{C}]$ as the largest element.

In this paper, the fuzzy entropy E is used as the fuzzy measure, and equation is as follows

$$E_p(U : c) = \frac{4}{c^P} \sum_{j=1}^{c} \sum_{i=1}^{n} u_{ij}(1 - u_{ij}) \tag{4}$$

in (4), c is the number of divided clusters. In addition, when $1 < c < n$, equation (4) must satisfy the following properties:

1) $0 \leqslant E_p(U : c) \leqslant \frac{4n(c-1)}{c^{p+1}}$
2) $E_p(U : c) = 0 \Leftrightarrow U$ is hard partition
3) $U = [\frac{1}{c}] \Leftrightarrow E_p(U : c) = \frac{4n(c-1)}{c^{p+1}}$

Proofs of these theorems have been given in reference [37], so they are omitted here.

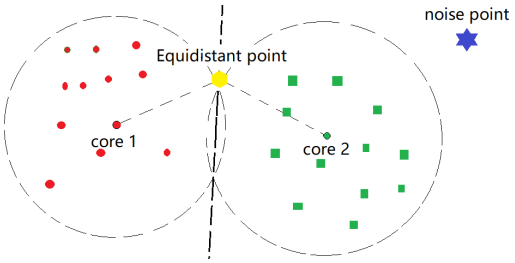

FIGURE 1. Illustation of the Spical points in this study.

What's more, Fan and Wu [38] proved that clustering with the use of fuzzy entropy as a fuzzy metric is feasible. The fuzzy entropy function in the reference [38] is a particular case where p is zero in the reference [37]. Meanwhile, the parameter p also plays an analytical role in clustering to overcome defects of partition coefficients and other similar clustering validity functions.

III. OPTIMIZATION METHOD

When the K-means algorithm is applicated in the processesing of data, there may occur special point problems such as equidistant points or noise, as shown in FIGURE 1. From equation (2), it can be known that the center point is controlled by the distance between points during iteration. So when there are special points, the clustering results of the K-means algorithm are very easily affected. Therefore, this paper proposes an improved algorithm based on fuzzy entropy (FMK-means), according to the characteristics of fuzzy entropywhich judges the set structure.

A. OPTIMIZATION DIRECTION

When processing data, the traditional K-means algorithm and the improved algorithm still focus on calculating the distance between points to get the globally optimal solution. However, their clustering effects on non-convex sets are not good. At the same time, as the initial cluster centers are randomly selected, the cluster center iteration is contracted evenly instead of conforming to the characteristic direction of data distribution. Hence, the problem of sensitivity to special points has not been solved [39]. When the membership degree is introduced into the fuzzy clustering FCM algorithm, although more information between points is referred to, the value of emphasis, m, related to the membership degree needs to be taken account of. There is still no ideal theoretical result for this problem [40]. None of the above clustering can assess the clustering effect by reflecting the set's overall ambiguity and information degree of the set.

Therefore, in order to solve the problem of intra-cluster measurement and initial center point, the overall feature space and the initial artificial cluster center point are needed in addition to the concept of fuzzy measurement. PEntropy usually functions in the description of the order of information: the higher the order of information is, the lower the fuzzy entropy will be, and vice versa. For example, reference [41], a simple fuzzy entropy constraint is added to the original FCM algorithm, which can suppress noise data to a certain extent. In reference [42], there emerged, based on the original FCM,

a membership equation with data distribution characteristics and the fuzzy clustering algorithm on entropy (FCMOE) with fuzzy entropy constraints. Both of them have a certain improving influence on the convergence direction of data. Reference [43], the ABC-kmeans algorithm is supplemented with the strategy of data features, performing better in clustering. Reference [44], based on the PSO-Kmeans algorithm, a two-step method is proposed by giving priority to a better initial clustering center, finally reducing the cost..

In summary, the introduction of the fuzzy entropy function makes fuzzy entropy the coefficient of the K-means algorithm's objective function. Meanwhile, an optimized clustering algorithm with Gaussian distribution can be derived by utilizing the distribution characteristics of the data to be clustered and taking fuzzy entropy as the cost function constraint, finally improving the anti-noise performance. Moreover, this algorithm gives a better initial center point to reduce iterations.

B. FMK-MEANS ALGORITHM DERIVATION

Plug fuzzy entropy into equation (1) as a measure of fuzzy degree., equation (5) can be obtained,in which the constraint condition can be got through the theorem 2.

$$SSE = \frac{4}{C^p}\sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}(1-u_{ij})(x_i-c_j)^2$$
$$s.t. \quad p = \log_c 4(c-1) - 1 \tag{5}$$

Concurrently, the distribution characteristic ω of the clustering function is introduced to make the newly added constraint condition contract according to it.

$$\omega = \frac{1}{nN}\sum_{i=1}^{n}\sum_{k=1}^{N}\frac{x_{ik}}{\max x_k} \tag{6}$$

in, N is the number of dimensions of elements in the cluster, and ω represents the normalized distribution characteristics of the clustered data set. According to the factor ω, let the fuzzy entropy meet the data distribution characteristics. The KKT condition of FMK-means algorithm with fuzzy entropy factor can be expressed as equation (7)

$$S = \frac{4}{C^p}\sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}(1-u_{ij})(x_i-c_j)^2 + \omega u_{ij}\ln u_{ij}$$
$$s.t. \quad g(u_{ij}) = u_{ij}\ln u_{ij} \leqslant 0$$
$$\omega \geqslant 0$$
$$\omega g(0) = \omega g(1) = 0 \tag{7}$$

Find the extremum here and define Lagrangian multiplication. A new function of (8) is available

$$L = \frac{4}{C^p}\sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}(1-u_{ij})(x_i-c_j)^2 + \omega u_{ij}\ln u_{ij}$$
$$+\sum_{i=1}^{n}\lambda_i(1-\sum_{j=1}^{c}u_{ij})$$
$$s.t. \quad \sum_{j=1}^{c}u_{ij} = 1 \tag{8}$$

Take the partial derivative of u_{ij} in (8) and make it zero. Through constraint (7), the degree of membership can be finally obtained, which can be expressed as equation (9)

$$u_{ij} = \frac{exp(-\frac{4(1-2u_{ij})(x_i-x_j)^2}{C^p\omega})}{\sum_{t=1}^{c} exp(-\frac{4(1-2u_{ij})(x_i-x_j)^2}{C^p\omega})} \tag{9}$$

Then calculate the partial derivative of the cluster center c_i from equation (8), and the cluster center can be expressed as equation (10)

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}(1-u_{ij})x_i}{\sum_{i=1}^{n} u_{ij}(1-u_{ij})} \tag{10}$$

C. INITIAL CENTER POINT SOLUTION OF FMK-MEANS ALGORITHM

In the traditional K-means algorithm, the initial center point is randomly allocated. If using noise or edge points as the initial point, a significant interference will definitely be caused to the result. To avoid similar situations, the initial point can be given artificially, and the initial point is required to be as close to actual point size as possible. According to the distribution of clustering centers, centers of different clusters must be arranged in a nearly linear way in one or more dimensions. An approximation effect can be achieved through the average value. Assume that the data set X includes n N-dimensional data divided into c clusters. Then the cluster center of the Ith dimension ($I = 1, 2, \cdots, N$) of any i cluster can be expressed as:

$$v_iI = \frac{3}{2}\frac{i}{c}\frac{1}{n}\sum_{j=1}^{n} x_{jI} \tag{11}$$

Get the initial center point $V_i = \{v_{i1}, v_{i2}, \cdots, v_{iI}, \cdots, v_{iN} | i = 1, 2, \cdots, c\}$

D. THE FLOW OF FMK-MEANS ALGORITHM

The traditional K-means algorithm randomly selects the initial point without considering distribution characteristics of actual data. It does not have soft clustering characteristics, which leads to the lack of robustness and stability of clustering results. Obviously, compared with the traditional K-means algorithm, it's more reasonable for the improved algorithm to work from multiple angles rather than a single one. In addition, the improved algorithm reflects a Gaussian distribution nature and can effectively overcome the problem of sensitivity to noise. The initial point can be given automatically, and a soft clustering algorithm with data characteristics is proposed in this paper. The specific steps are as follows:

Step 1 For a given data set, the mean points of all samples are calculated by referring to equation (11). The mean point is taken as the first clustering center, denoted as$V^{(0)}$. The distribution characteristic of the overall data, ω, is calculated through equation (6).

Step 2 Calculate the $u_{ij}^{(0)}$ of all samples, and the membership degrees will be determined according to equation (9). Similarly, at the beginning, all the sample points are initialized, which means all of the membership degrees should be zero.

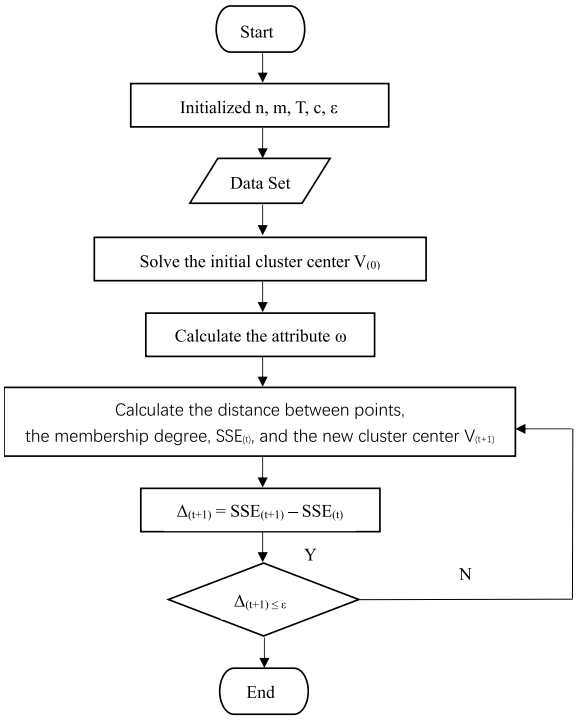

FIGURE 2. Algorithm flow chart.

Step 3 Conduct the algorithm FMK-means calculation on given data sets. Constantly iterate u_{ij} and c_j, until the termination condition of the iteration is reached.

Step 4 Output final clustering results.

IV. FMK-MEANS ALGORITHM COMPLEXITY

In terms of time cost, the time complexity of the traditional K-means algorithm is $O(i \times c \times n \times m)$, where c is the number of clusters, i is the number of iterations required for convergence, and n is the number of the the number of the data set contains, m is the number of attributes of each data. If the convergence occurs faster, the value of i will usually relatively small. As long as the number of clusters c is significantly less than n, the K-means algorithm is linearly related to n.

In the optimization algorithm, the time complexity of the algorithm to solve the initial center point can be expressed as $O(n \times m)$. The time complexity of solving the membership degree in a single iteration in the subsequent iteration process is $O(c \times n \times m)$, and the time complexity of cluster centers in each iteration is $O(c \times n \times m)$. By introducing the above data to equation (5), the time cost of the FMK-means algorithm is $O(c \times n \times m)$. In summary, the total cost of each iterations is $2O(c \times n \times m)$. Assuming that the number of iterations is i, then the total cost time complexity is $O(i \times c \times n \times m)$. Finally, the time complexity of the FMK-means algorithm can be determined to be $O(kn)$.

In reality, in consideration of the actual effects of the initial center points, the number of iterations of the improved algorithm will be much smaller than the number of iterations i of the traditional K-means algorithm.

Algorithm 1 FMK-Means (Fuzzy Metrics K-Means)

Require: Data set, the Maximum number of iterations T, Threshold ε, the Number of clusters c

Ensure: Clustering results of the data set

```
initialize the Array List;
compute V_j (D);
    FOR(each sample j∈K){
        FOR(each sample I∈N){
        compute v_jI;
        }
    }
select Center V_j;
PRINTF(K,Initial Center V);
WHILE(data sets D!=null){
compute u_ij(D);
    FOR(each sample jinK){
        FOR(each sample iinD){
        compute u_ij;
        }
    }
select Membership u_ij;
PRINTF(D,Initial Membership u_ij);
FMK-means inputer(D, K,Initial Center V, D,Initial Membership; u_ij)
WHILE(new center!=original center){
    FOR(each center V_j ∈ V) {
        FOR(each center V_j ∈ V){
            compute SSE(u_ik,d_ik);
        }
        IF(SSE(u_ik,d_ik)=Min SSE(u_ij,d_ij)){
            Center V_k <- sample I;
        }
    } //END FOR;
    compute NEW Center Vj =
    Mean(sample(i &&(i ∈ ClusterV_j)));
} // END WHILE;
PRINTF(Cluster V_j);
```

V. SIMULATION EXPERIMENT OF CLUSTERING

A. EXPERIMENTAL ENVIRONMENT AND DESIGN

The experiment was conducted with the system of Windows 10, 16GB physical memory, CPU frequency of 3.20GHz and the Python 3.8 platform. Comparisons were realized between the FMK-means algorithm, K-means algorithm, K-means++ algorithm, FCM algorithm,Alternative Fuzzy k-means (AFKM) algorithm [45] and FCMOE algorithm. The k-means algorithm, k-means++ algorithm, and FCM algorithm were implemented by calling the Scikit-learn [46], [47]. The experimental data were 5 data sets (iris, balance, phoneme, ring and HTRU) [48]–[52], and detailed parameters are shown in TABLE 1. Since the reference data has a certain degree of the reference data, the clustering results have a more intuitive comparison effect.

TABLE 1. Attribution of data.

Data	Number	Attributes	categories
Iris	150	4	3
Balance	625	4	3
Phoneme	5404	5	2
Ring	7400	20	2
HTRU	17898	9	2

TABLE 2. Performance comparison of Iris clustering algorithm.

algorithm	result	frequency	Accuracy(%)	time(ms)
K-means	78.94	11	84.0	23
K-means++	76.69	18	84.1	33
FCM	285.14	57	86.7	37
AFKM	348.26	41	89.7	34
FCOME	329.64	27	90.7	46
FMK-means	326.88	21	91.1	32

TABLE 3. Performance comparison of balabce clustering algorithm.

algorithm	result	frequency	Accuracy(%)	time(ms)
K-means	3474	19	59.7	62
K-means++	3470	24	61.2	80
FCM	1723	521	76.7	984
AFKM	1696	492	77.1	1208
FCOME	1655	296	78.9	613
FMK-means	1577	297	79.2	633

TABLE 4. Performance comparison of phoneme clustering algorithm.

algorithm	result	frequency	Accuracy(%)	time(ms)
K-means	12837.80	19	74.1	102
K-means++	12837.87	19	74.8	103
FCM	9832.97	32	67.7	88
AFKM	9653.47	30	70.1	96
FCOME	9322.71	23	78.1	88
FMK-means	9297.87	21	78.2	84

The parameter are $t = 10000$, and the threshold $\varepsilon = 0.0001$. Becauseit is pointed out in reference [53] that when the fuzzy index is $1.5 - 2$, the clustering effect is ideal, so set $m = 2$ in FCM.

In order to verify the effect of the improved FMK-means algorithm, two sets of comparative experiments were designed. Experiment 1 compared the above data's clustering results with different clustering algorithms to verifythe increase of the improved algorithm on the clustering effects of the traditional algorithm. In the experiment, each group of algorithms took the average of ten calculation results for different data groups in turn as the clustering results. Experiment 2 verified the anti-noise ability of the improved algorithm and the traditional algorithm by adding noise data to the sample data.

B. EXPERIMENTAL RESULTS AND ANALYSIS

1) ALGORITHM CLUSTERING RESULTS

The clustering performance comparison of the above six algorithms is shown in the table below

In the above results, we found that the amount of data and dimensions have different effects on different ranges of the same method. In addition, in the same data set, the output results of different algorithms were also quite different. Based on the data in TABLE 2 to TABLE 6, set the

TABLE 5. Performance comparison of ring clustering algorithm.

algorithm	result	frequency	Accuracy(%)	time(ms)
K-means	319809084566	26	64.4	172
K-means++	319808889384	22	64.3	159
FCM	167235493497	7	83.9	86
AFKN	164663234143	8	84.3	93
FCOME	130706814338	5	86.9	67
FMK-means	126740804128	5	87.5	37

TABLE 6. Performance comparison of HTRU clustering algorithm.

algorithm	result	frequency	Accuracy(%)	time(ms)
K-means	122775535	186	87.0	288
K-means++	122771113	133	87.1	303
FCM	94000261	57	73.8	547
AFKN	93799064	51	74.2	624
FCOME	87437421	49	89.1	496
FMK-means	87547142	48	88.3	499

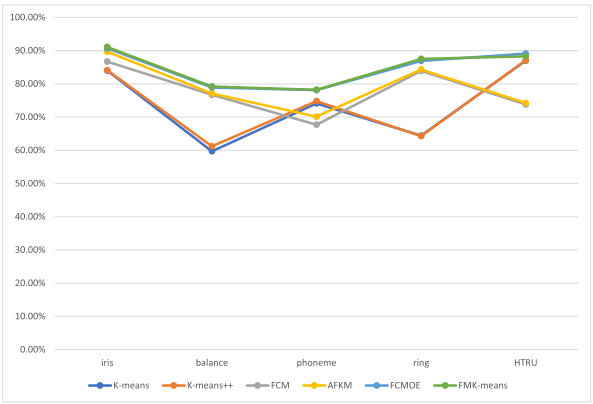

FIGURE 3. Comparison of clustering accuracy of the six algo-Rithms.

results of different algorithms in the five cases of 500, 2000, 20000,100000(8) and 100000(19).

It can be seen from the accuracy line chart of FIGURE 3 that the six clustering algorithms perform best on Iris. Hard clustering performs the worst on the accuracy of Balance data, because the banlance data is not a continuous variable. The worst effect on soft clustering in phoneme is due to the fact that different types of data are more scattered. However, considering data distribution characteristics, the two algorithms FCMOE and FMK-means stand out for their high accuracy. It also somewhat reflects that purely measuring the distance between points in calculation is unreasonable.

It is not difficult to find from the comparison that FMK-means in this paper has several indicators ranking the best in the performance test. In several experimental data sets, the FMK-means algorithm has the highest average accuracy and is close to the FCMOE algorithm. The reason is that both of them offer the initial cluster center, avoid the problem of random initial points, and also take into account fuzzy entropy measurement and the overall distribution characteristics of data, hence more comprehensive than other algorithms. Although changing the metric in the AFKM algorithm improves the robustness of the AE metric, its clustering effect is not significantly improved compared to the FCM.

To summarize, for low-dimensional small data clustering, when the stopping condition is satisfied, the FMK-means

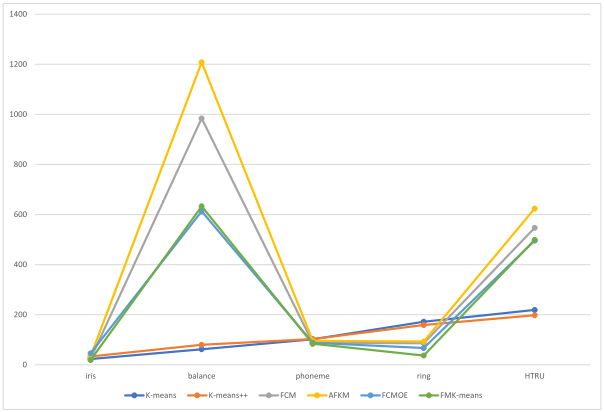

FIGURE 4. Clustering time of the six algo-Rithms.

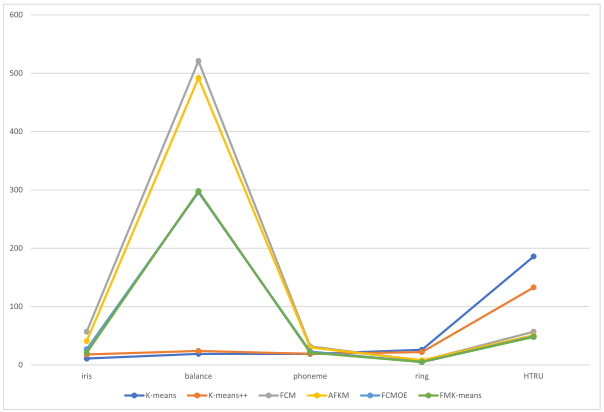

FIGURE 5. Number of iterations of the six algo-Rithms.

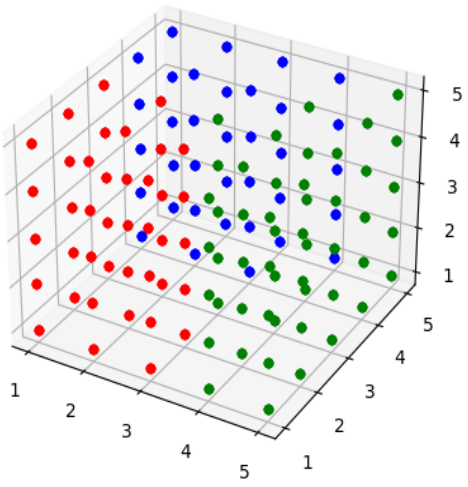

FIGURE 6. Data distribution of banlance.

TABLE 7. Initial cluster center comparison.

	Iris	Phoneme	Ring	HTUR
The Random	27	25	7	54
Equation (11)	21	21	5	48

TABLE 8. Anti-noise performance.

	10%	15%	20%
Error rate of K-mean	32.5%	35.7%	39.4%
Error rate of K-means++	32.3%	35.7%	39.0%
Error rate of FCM	33.0%	36.6%	40.1%
Error rate of AFKM	31.3%	34.5%	38.2%
Error rate of FCMOE	8.9%	10.1%	12.2%
Error rate of FMK-means	8.7%	9.2%	10.1%

algorithm has the fewest iterations and the highest accuracy rate. In the case of rising data volume from Iris to Phoneme, the number of iterations of the traditional algorithm increases at an incredible speed, but the soft clustering methods have fewer iterations and the time cost changes smoothly. The main reason is that the degree of membership can better reflect the relationship between multiple points rather than simply the distance between two points, thereby speeding up the convergence. Moreover, the algorithms of FMK-means and FCMOE have a smaller number of iterations due to the artificial initial center. Besides, when dimensions increase from Phoneme to Ring, the accuracy of K-means algorithms decreases, but the accuracy of fuzzy clustering does not experience a similar change and clustering algorithms with fuzzy entropy as the constraint posess the highest accuracy. Since fuzzy entropy describes the nature of the overall fuzzy degree, and since the more specific the set is, the lower fuzzy entropy is, FMK-means has a higher accuracy and a lower clustering output. In the Balance data, there appear abnormal feedback results in the above several algorithms because the balance data is not convex data, and the distribution is shown in FIGURE 6. Therefore, in spite of a small data volume of the sample, the accuracy of the traditional K-means algorithm is not high. The FMK-means and the FCMOE algorithm, though taking a little bit longer time, can gain a high accuracy due to the use of data distribution characteristics as the contraction direction. Both Ring and HTRU are at the level of 100,000, but the impact of their dimensionality on all algorithms is higher than that of quantity.

Even though the FCMOE requires fewer iterations than the traditional K-means algorithm and FCM algorithm, there is still a certain gap on accuracy and convergence results between it and the FMK-means algorithm. The main reason is that fuzzy entropy is used as the coefficient in the optimization algorithm. Although both introduce an adjustment factor that indicates the distribution characteristics of the data set, it is more reliable in the reduction direction. However, due to the calculation with fuzzy entropy introduced, the output of the FMK-means algorithm is the smallest, and the result of the overall distribution is more unambiguous.

2) ANTI-NOISE ABILITY COMPARISON

For the purpose of comparing the anti-noise ability of the FMK-means algorithm and the other five algorithms, 10%, 15% and 20% of the noise data are added to the Iris data. Then use these algorithms to complete the clustering and check the anti-noise ability. The experimental results are shown in TABLE 8.

According to the experimental results, the FMK-means algorithm is less affected by noise and has a good anti-noise performance. Compared with traditional K-means algorithms, FMK-means, with artificial initial cluster centers,

effectively avoids the interference of randomly selected special points, reduces iterations and possesses some anti-noise ability. Fuzzy entropy is introduced as a new constraint in the FMK-means algorithm. Secondly, in the FMK-means algorithm, data distribution characteristics are also introduced into the membership degree, so that the membership function is supplemented with Gaussian distribution characteristics. By doing so, it ensures that the reducing direction of entropy conforms to that of distribution characteristics during the algorithm convergence, thereby reducing the impact of non-convex clustering.

Comparison between the clustering effects of different algorithms and the analysis the clustering results make it clear that the improved FMK-means algorithm has stronger advantages in terms of convergence speed, iterations and clustering accuracy.

VI. CONCLUSION

This paper proposes an improved K-means algorithm based on fuzzy entropy. Fuzzy entropy characterized in its description of the set's fuzzy degree contributes to the solution of the problem that the traditional K-means algorithm is extremely sensitive to special points. Meanwhile, a new solution to the initial center point is proposed, which avoids the defect that the initial point is randomly selected and the risks of a special point or a local optimal solution. In the optimization algorithm, the factor representing data set distribution characteristics is specified for fuzzy entropy as ω, so that the clustering result solution has Gaussian distribution characteristics, which improves the accuracy and noise resistance of the clustering algorithm. However, the processing of multi-dimensional data, such as text data and image data, is not covered in this paper. Further research tasks are listed as how to select a correct way to reduce dimensionality and how to reasonably introduce other types of ambiguity.

REFERENCES

[1] Y. Wu, "General overview on clustering algorithms," *Comput. Sci.*, vol. 42, pp. 491–499,524, Jun. 2015.
[2] Y. Jing and J. Wang, "Tag clustering algorithm LMMSK: Improved K-means algorithm based on latent semantic analysis," *J. Syst. Eng. Electron.*, vol. 28, no. 2, pp. 374–384, Apr. 2017.
[3] S. Sambasivam and N. Theodosopoulos, "Advanced data clustering methods of mining Web documents," *Issues Informing Sci. Inf. Technol.*, vol. 3, pp. 563–579, Jan. 2006.
[4] J. Xu, K. Zheng, M.-M. Chi, Y.-Y. Zhu, X.-H. Yu, and X.-F. Zhou, "Trajectory big data: Data, applications and techniques," *J. Commun.*, vol. 36, no. 12, pp. 97–105, 2015.
[5] *IDC*. Accessed: Oct. 31, 2020. [Online]. Available: https://www. idc.com/
[6] P. M. Sosa, G. S. Carrazoni, R. Gonçalves, and P. B. Mello-Carpes, "Use of facebook groups as a strategy for continuum involvement of students with physiology after finishing a physiology course," *Adv. Physiol. Edu.*, vol. 44, no. 3, pp. 358–361, Sep. 2020.
[7] F. Yu, K. Peng, and X. Zheng, "Big data and psychology in China," *Chin. Sci. Bull.*, vol. 60, nos. 5–6, pp. 520–533, Feb. 2015.
[8] K. Jahanbin and V. Rahmanian, "Using Twitter and Web news mining to predict COVID-19 outbreak," *Asian Pacific J. Tropical Med.*, vol. 13, pp. 378–380, Jul. 2020.
[9] J.-G. Sun, "Clustering algorithms research," *J. Softw.*, vol. 19, no. 1, pp. 48–61, Jun. 2008.
[10] H. Dai, Z. Chang, and N. Yu, "Understanding data mining," in *Introduction to Data Mining*, 1st ed. Beijing, China: Tsinghua Univ. Press, 2015, pp. 1–27.
[11] Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining," *DMKD*, vol. 3, no. 8, pp. 34–39, 1997.
[12] Z. Huang, "Extensions to the K-means algorithm for clustering large data sets with categorical values," *Data Mining Knowl. Discovery*, vol. 2, no. 3, pp. 283–304, Sep. 1998.
[13] C. Ding and X. He, "K-nearest-neighbor consistency in data clustering: incorporating local information into global optimization," in *Proc. ACM Symp. Appl. Comput.*, Nicosia, 2004, pp. 584–589.
[14] G. Zhang, C. Zhang, and H. Zhang, "Improved K-means algorithm based on density canopy," *Knowl.-Based Syst.*, vol. 145, pp. 289–297, Apr. 2018.
[15] Z. Xianchao, "Cluster-based partition algorithm," in *Data Clustering*, 3rd ed. Beijing, China: Science Press, 2017, pp. 37–62.
[16] D. Yueming, W. Minghui, Z. Ming, and W. Yan, "Optimizing initial cluster Centroids by SVD in K-means algorithm for Chinese text clustering," *J. Syst. Simul.*, vol. 30, pp. 244–251, Oct. 2018.
[17] L. T. Z. Can, "Nearest neighbor optimization k-means clustering algorithm," *Comput. Sci.*, vol. 46, pp. 216–219, Nov. 2019.
[18] W. L. L. X. Z. Liangjun, "Research on K-means clustering algorithm based on ARIA," *J. Sichuan Univ. Sci. Eng. (Natural Sci. Ed.)*, vol. 32, pp. 65–70, Apr. 2019.
[19] Y. Xiaoli, "Design of center random drift (CRD) K-means clustering algorithm," *J. Changchun Univ.*, vol. 27, pp. 35–38, Aug. 2017.
[20] C. Guo and Y. Zang, "Clustering algorithm based on density function and nichePSO," *J. Syst. Eng. Electron.*, vol. 23, no. 3, pp. 445–452, Jun. 2012.
[21] Y. Jinping, Z. Jie, and M. Hongbiao, "K-means clustering algorithm based on improved artificial bee colony algorithm," *J. Comput. Appl.*, vol. 34, pp. 1065–1069, Jun. 2014.
[22] L. Jia-Ming, K. Li-Qun, and Y. Hong-Hong, "K-means clustering algorithm optimized by Gray Wolf," *Chin. Sciencepaper*, vol. 14, pp. 778–782 and 807, Aug. 2019.
[23] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Comput. Geosci.*, vol. 10, nos. 2–3, pp. 191–203, Jan. 1984.
[24] Y. S. Song and D. C. Park, "Fuzzy C-means algorithm with divergence-based kernel," in *Proc. Int. Conf. Fuzzy Syst. Knowl. Discovery*, Berlin,Germany, 2006, pp. 99–108.
[25] C. L. Chowdhary and D. P. Acharjya, "Clustering algorithm in possibilistic exponential fuzzy C-Mean segmenting medical images," *J. Biomimetics, Biomater. Biomed. Eng.*, vol. 30, pp. 12–23, Jan. 2017.
[26] C. Xia and F. Fang, "Fuzzy Clustering Methods," *Life Sci. Instrum.*, vol. 11, pp. 33–37, Dec. 2013.
[27] W. Chengmao and S. Jiamei, "Adaptive robust picture fuzzy clustering segmentation algorithm," *J. Huazhong Univ. Sci. Technol. (Natural Sci. Ed.)*, vol. 47, pp. 115–120, Sep. 2019.
[28] C. L. Chowdhary and D. P. Acharjya, "A hybrid scheme for breast cancer detection using intuitionistic fuzzy rough set technique," *Int. J. Healthcare Inf. Syst. Informat.*, vol. 11, no. 2, pp. 38–61, Apr. 2016.
[29] V. Cherkassky and F. Mulier, "Support Vector Machines," in *Learning from Data Concepts, Theory, and Methods*, vol. 1, 9th ed. New York, NY, USA: Wiley, 2007, pp. 413–417.
[30] J. Li, X. B. Gao, and L. C. Jiao, "A new feature weighted fuzzy clustering algorithm," *Acta Electronica Sinica*, vol. 34, pp. 89–92, Jan. 2006.
[31] C. Palanisamy and S. Selvan, "Efficient subspace clustering for higher dimensional data using fuzzy entropy," *J. Syst. Sci. Syst. Eng.*, vol. 18, no. 1, pp. 95–110, Mar. 2009.
[32] C. L. Chowdhary and D. P. Acharjya, "Breast cancer detection using intuitionistic fuzzy histogram hyperbolization and possibilitic fuzzy C-mean clustering algorithms with texture feature based classification on mammography images," in *Proc. Int. Conf. Adv. Inf. Commun. Technol. Comput.*, 16th ed. Bikaner, India: ACM Press, 2016, pp. 1–6.
[33] V. Dey, D. K. Pratihar, and G. L. Datta, "Genetic algorithm-tuned entropy-based fuzzy C-means algorithm for obtaining distinct and compact clusters," *Fuzzy Optim. Decis. Making*, vol. 10, no. 2, pp. 153–166, Jun. 2011.
[34] C. L. Chowdhary and D. P. Acharjya, "Segmentation of mammograms using a novel intuitionistic possibilistic fuzzy C-mean clustering algorithm," *Nature Inspired Comput.*, vol. 652, pp. 75–82, Jan. 2018.
[35] H. T. Hong and Yonghong, "Automatic pattern recognition of ECG signals using entropy-based adaptive dimensionality reduction and clustering," *Appl. Soft Comput.*, vol. 55, pp. 238–252, Jun. 2017.
[36] J. Zejun, "Fuzzy Set," in *Theory and Methods of Fuzzy Mathematics*. Beijing, China: Publishing House of Electronics Industry, 2015, pp. 23–27.
[37] M. X. QING and SUN, "A new clustering effectiveness function: Fuzzy entropy of fuzzy partition," *CAAI Trans. Intell. Syst.*, vol. 10, pp. 75–80, Jan. 2015.

[38] F. Jiulun and W. Chengmao, "Clustering validity function based on fuzzy entropy," *Pattern Recognit. Artif. Intell.*, vol. 14, no. 4, pp. 390–394, Dec. 2001.
[39] Z. Liu and Y. Hu, "Research on FCM clustering optimization algorithm for self-adaptive bacterial foraging," *Mod. Electron. Technique*, vol. 43, pp. 144–148, Mar. 2020.
[40] X. J. Gao and Pei, "A study of weighting exponent m in a fuzzy C-means algorithm," *Acta Electronica Sinica*, vol. 28, pp. 80–83, Apr. 2000.
[41] T. Chaira, "A novel intuitionistic fuzzy c means clustering algorithm and its application to medical images," *Appl. Soft Comput.*, vol. 11, no. 2, pp. 1711–1717, Mar. 2011.
[42] S. Liao, J. Zhang, and A. Liu, "Fuzzy C-means clustering algorithm by using fuzzy entropy constraint," *J. Chin. Comput. Syst.*, vol. 35, pp. 189–193, Feb. 2014.
[43] U. H. Atasever, "A novel unsupervised change detection approach based on reconstruction independent component analysis and ABC-kmeans clustering for environmental monitoring," *Environ. Monitor. Assessment*, vol. 191, no. 7, pp. 1–11, Jun. 2019.
[44] C. Ibrahim and I. Mougharbel, "Two stages K-means and PSO-based method for optimal allocation of multiple parallel DRPs application & deployment," *IET Smart Grid*, vol. 3, no. 2, pp. 216–225, May 2019.
[45] T. Hu, "Discussion of improving fuzzy K-means clustering," M.S. thesis, Dept. School Data Comput. Sci., Sun Yat-Sen Univ., Guangzhou, China,2010.
[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[47] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt and G. Varoquaux, "API design for machine learning software: Experiences from the scikit-learn project," in *Proc. ECML PKDD Workshop, Lang. Data Mining Mach. Learn.*, 2013, pp. 108–122.
[48] *Iris Uci*. Accessed: Oct. 31, 2020. [Online]. Available:http://archive.ics.uci.edu/ml/datasets/Iris
[49] *Balance Uci*. Accessed: Oct. 31, 2020. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Balance+Scale
[50] *phoneme DataHub* Accessed:Oct. 31.2020. [Online]. Available: https://datahub.io/machine-learning/phoneme
[51] *Ring KEEL*. Accessed: Oct. 31, 2020. [Online]. Available: https://sci2s.ugr.es/keel/datasets.php
[52] R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, and J. D. Knowles, "Fifty years of pulsar candidate selection: From simple filters to a new principled real-time classification approach," *Monthly Notices Roy. Astronomical Soc.*, vol. 459, no. 1, pp. 1104–1123, Apr. 2016.
[53] N. R. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model," *IEEE Trans. Fuzzy Syst.*, vol. 3, no. 3, pp. 370–379, Aug. 1995.

XINYU GENG is currently a Professor with the School of Computer Science, Southwest Petroleum University, Chengdu, China. His main research interests include data mining and artificial neural networks.

YUKUN MU received the bachelor's degree in network engineering from Southwest Petroleum University, Chengdu, China, where he is currently pursuing the M.S. degree in computer science and technology. His research interests include fuzzy mathematics and data mining.

SENLIN MAO received the bachelor's degree in city underground space engineering from Southwest Petroleum University, Chengdu, China, where he is currently pursuing the M.S. degree in computer science and technology. His research interests include grey system theory, data mining, and artificial neural networks.

JINCHI YE received the bachelor's degree in information management and information system from Southwest Petroleum University, Chengdu, China, where he is currently pursuing the M.S. degree in computer science and technology. His research interests include grey system theory, data mining, and artificial neural networks.

LIPING ZHU received the bachelor's degree in computer science and technology from Sichuan Normal University, Chengdu, China, where she is currently pursuing the M.S. degree in engineering management. Her research interests include data mining and artificial neural networks.