Reserch Article

Gaussian Kernel Fuzzy C-Means Algorithm for Service Resource Allocation

Wei Jiang,1 Xi Fang,2 and Jianmei Ding3

1College of Economics and Management, Xi’an University of Posts and Telecommunications, Xi’an 710061, China
2School of Economics and Management, Shanghai Institute of Technology, Shanghai 200235, China
3School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China

Correspondence should be addressed to Xi Fang; fangxi@sit.edu.cn

Received 30 June 2020; Revised 20 November 2020; Accepted 5 December 2020; Published 19 December 2020

1.Introduction

Clustering is an unsupervised learning method that is not reliant on predefined classes and training datasets with class labels. Clustering objects are divided into classes or clusters on the basis of feature similarity measurement. Hence, the same clusters share high similarities within the same cluster but largely differ from each other between different clusters. Traditional clustering methods are primarily based on the partition, hierarchy, grid, density, and model. As data mining in rapid development necessitates higher requirements for clustering, a clustering algorithm based on sample attribution, preprocessing, similarity measurement, allocation and scheduling, update strategy, and measurement [1, 2] have been advanced and applied to data mining [3, 4]. Considering the fuzziness of membership between sample points and cluster centers, the objective function-based fuzzy c-means (FCM) algorithm still prevails in theory and practice.

The core of an FCM algorithm is to design and determine a clustering center. The design mainly consists of quantifying cluster centers, locating them, and scheming an objective function accordingly. The cluster centers are quantified manually in most cases, or their optimal amount is determined in a given range using information entropy and other methods. For example, Duan and Wang [5] indicated that the clustering center was acquired by multiattribute information with broken-line fuzzy numbers. A novel clustering algorithm, Nei Mu, was proposed in [6] based on which datasets are converted into data points of attribute space to construct a directed graph of K-nearest neighbors. This algorithm contributes to upgrading the clustering of data with large density fluctuation and an arbitrary distribution, but not all data points have K-nearest neighbors. Xue and Sha [7] initiated a coordinate-based density method using a gray prediction model of a clustering algorithm to determine the initial clustering center.

A clustering center should be determined with modifications in a dynamic process. The existing determination methods largely include K-means clustering algorithms, partition- and density-based clustering algorithms, clustering algorithms based on the local density of data points, and KZZ algorithms. For these, the K-means algorithm is used with a given initial center, whereas partition- and density-based clustering algorithms are used to determine the initial clustering center by a density function of sample...
points using max-min distance means or the maximum distance product method. Zhang and Wang [8] pointed out the nearest data points were bracketed to facilitate location of other clustering centers at the same time that a resolution with high constraints was added to the objective function. Chiu Stephen [9] defined measures for each data point to identify the initial clustering center. Agustin et al. [10] studied a group genetic algorithm, aiming to improve the performance of group clustering by coding and defining fitness functions. A semisupervised clustering algorithm was put forward in [11] via the kernel FCM clustering algorithm with clustering errors containing labeled and unlabeled data to design the objective function. Since FCM fails to deal with noise, an efficient kernel-induced FCM based on a Gaussian function was presented in [12] to improve the objective function.

The following cases are some of the existing FCM studies. Qian and Yao [13] focused on high sensitivity to the initial center point and introduced three incremental fuzzy clustering algorithms for large-scale sparse high-dimensional datasets. Niu and She [14] proposed a fast parallel clustering algorithm based on cluster initialization. By generating a hierarchical K-means clustering tree to auto-select the number of clusters, Hu [15] obtained better clustering results. Aiming at the high time complexity of traditional FCM algorithms, a single-pass Bayesian fuzzy clustering algorithm was advocated for large-scale data in [16], which boosted its performance in time complexity and convergence. Zhou et al. [17] introduced the neighborhood information of multidimensional data to improve the clustering algorithm, increasing the robustness of outliers and noise points. Chen and Liu [18] designed a clustering algorithm on the minimum connected dominating set to remedy the defect that common algorithms easily fall into local minimum points. Xie et al. [19] combined the GWO algorithm with the principle of maximum entropy in a multidimensional big data environment. Duan and Wang [5] described multiple attributes of the objects to be clustered as polygonal fuzzy numbers, and a clustering algorithm was designed accordingly. By advancing an adaptive algorithm for the entropy weight of the feature weight of FCM, Huang et al. [20] focused on the influence of the feature weight on a clustering algorithm. Taking the preference vectors’ clustering degree as a neighborhood similarity, Xu and Fan [21] aimed at constructing a heuristic clustering algorithm for multiattribute complex large group clustering and decision.

These documents focus on FCM algorithm-associated issues, but few achievements have been made in big data scenarios. In view of the differences between large data point clustering and small sample point clustering, the sample points of big data were simplified in this paper, making FCM more applicable for big data scenarios. Next, an FCM algorithm was designed by taking both long between-class distances and short inner-class distances into consideration, which traditional FCM algorithms failed to do. This study thus provides theoretical and practical guidance for data clustering in a big data environment.

2. Gaussian Kernel FCM Clustering Algorithm

Since service resources are generally allocated in multiple ways, and there is a reciprocal relationship between the limited resources in one channel of allocation and those in another in terms of resource quantity, group consistency is beyond reach in which different resource consumers prefer different channels, leading to changing evaluation data. If the price mechanism fails to optimize the service resource allocation, consumer demands should be considered while pursuing social benefits to attain higher efficiency of resource allocation. Consumers primarily feature heterogeneity, conflicts of interest, and differences in evaluation forms, which necessitate decomposition of the customer group to divide the large-scale consumer groups into several small clusters, thus simplifying resource coordination.

Suppose that the consumer subject of a service resource is expressed as \( X = (x_1, \ldots, x_n) \), individual consumer as \( x_i, i \in (1, \ldots, n) \), number of channels (data dimension) as \( p \), evaluation data as \( x_{ij}, j \in (1, \ldots, p) \), \( u_{ik} \) is the membership of sample \( i \) in Class \( k \), with fuzzy matrix \( U = [u_{ik}] \) provided that there are \( c \) classes, and \( y_k, k \in (1, \ldots, c) \) represents cluster centers. Its objective function can be represented by the Gaussian kernel FCM clustering algorithm [8]:

\[
J(U, V) = 2 \sum_{i=1}^{n} \sum_{k=1}^{c} u_{ik}^m (\beta - \phi(x_i, y_k)),
\]

where \( \phi(x_i, y_k) = \exp(-\|x_i - y_k\|^2/\sigma^2) \), \( \beta \) is a characteristic constant of the Gaussian function, and \( m \) is a fuzzy index used to control the fuzzy degree of classification. The higher the index, the higher the fuzzy degree. \( \sigma^2 \) is the variance of the given data. Hence,

\[
u_{ik} = \frac{1}{\sum_{k=1}^{c} (1/(\beta - \phi(x_i, y_k)))^{(1/m - 1)}} \times \frac{1/(\beta - \phi(x_i, y_k))^{(1/(m-1))}}{\sum_{i=1}^{n} u_{ik}^m \beta \phi(x_i, y_k)},
\]

\[
y_k = \frac{\sum_{i=1}^{n} u_{ik}^m \beta \phi(x_i, y_k) x_i}{\sum_{i=1}^{n} u_{ik}^m \beta \phi(x_i, y_k)}.
\]

If \( \|v_{\text{present}} - v_{\text{previous}}\| \leq \epsilon \), then the iteration is discontinued, at which there is optimum classification. Both traditional FCM algorithms and the Gaussian kernel FCM clustering algorithm focus on the inner-class distance instead of between-class distance. Result-oriented, both values should be considered in order to obtain better clustering. Due to the large number of consumers to whom service resources are allocated, direct computing of membership will lead to problems such as high computing complexity and slow convergence of the optimal solution, resulting in a...
decline in clustering efficiency. Therefore, preprocessing of data points should occur prior to clustering in order to reduce the number of data points that need clustering and enhance the scalability of the clustering algorithm.

3. Preprocessing of Evaluation Information of Consumers

∀x_i, x_j can be considered as a constraint data pair. The Euclidean distance formula is deployed to calculate its distance:

\[ d(x_i, x_j) = \|x_i - x_j\| = \sqrt{\sum_{p=1}^{n}(x_{ip} - x_{jp})^2}. \]

(4)

Set \( \varepsilon \) and \( \gamma \) in advance for \( d(x_i, x_j) \) (both \( \varepsilon \) and \( \gamma \) can take lower values for more accurate classification).

(i) If \( d(x_i, x_j) \leq \varepsilon \), then it is considered that \( x_i \) and \( x_j \) are extremely close and can be placed into one class
(ii) If \( d(x_i, x_j) \geq \gamma \), it is considered that \( x_i \) is far from \( x_j \), and bracketing them together is next to impossible
(iii) Data points between \( \varepsilon \) and \( \gamma \) cannot be effectively identified

To delete data points quickly, the characteristics of distances between data points and the possibility of clustering different data points should be considered and investigated in the preprocessing procedure. Deletion should be done via the following steps:

(i) Step 1: take data points \( x_i \) and \( x_j \) with the smallest distance in \( X \) to meet \( d(x_i, x_j) \leq \varepsilon \), and combine \( F_1 = \phi \) with \( x_i \) and \( x_j \). Then, \( F_1 = \{x_i, x_j\} \rightarrow F_1 \), and \( X = \{x_i, x_j\} \rightarrow X \).

(ii) Step 2: take the mean value of \( x_i \) and \( x_j \) by \( (x_i + x_j)/2 \) as a new data point, and identify data point \( x_t \) in \( X \) whose mean value is less than \( \varepsilon \). Then, \( F_1 = \{x_t\} \rightarrow F_1 \) and \( X = \{x_t\} \rightarrow X \).

(iii) Step 3: take \( \{x_i, x_j, x_t\} \) as a new data point with a value of \( ((x_i + x_j)/2 + x_t)/2 \). Repeat Step 2 until no new data points can be found, and form new sets \( F_1 \) and \( X \).

(iv) Step 4: repeat Steps 1 to 3 for Set \( X \) form Set \( Y \), including \( F_1, \ldots, F_m \) and \( X \), wherein the final mean values of data points in \( F_1, \ldots, F_m \) are taken as new data points, respectively.

(v) Step 5: let data points in \( Y \) be nodes in the graph based on graph theory, and the connecting line of nodes be the distance. If the distance is greater than \( \gamma \), then the connecting line is deleted, thus forming a connected network graph. Assuming that points \( a, b, \) and \( c \) make a circle, and \( a \) is the farthest from \( b \), it can be considered that there is a higher probability of forming a cluster by \( a \) and \( b \) than by \( a \) and \( c \), or \( b \) and \( c \), so the connecting line can be deleted. Here, Graph \( G \) without cycles is the connected network graph.

(vi) Step 6: in Graph \( G \) with a plurality of nodes, the nodes are sorted by the number of adjacent points. Each node has a plurality of nondominated points to make a cluster.

(vii) Step 7: since cluster sides vary in length, it is difficult to generate an effective cluster set for clusters with their average value as its sides. Deletion leads to basically equal distance from each data point in a cluster to the center of the cluster. Thus, point estimation may be adopted to work out its expected value.

Given that Cluster \( C_i \) has \( q \) neighbors, its sample variance is \( s_i \), \( (x_i - \bar{x}/(s_i/\sqrt{q})) \leq t_{1-\alpha} \). An adaptive k-nearest neighbor algorithm is used to search for a data point closest to or within a set distance from the given data point and merge it into clustering components for clustering fusion. Therefore, \( x_i \) that satisfies the formula will be included in Cluster \( C_i \); otherwise, the data point is deleted from \( C_i \). If included in multiple classes, the data point will enter the cluster as a matter of priority, where it is the nearest to the cluster center, so the average value of all data points in Cluster \( C_i \) is data point \( z_i \).

(viii) Step 8: data points are downsized using the approach mentioned above, and the pertinence of clustering by Set \( Z \) is strengthened. The original dataset \( X \) becomes Set \( Z = (z_1, \ldots, z_m) \).

4. Clustering Algorithm of Consumer Evaluation

The number of clusters and the initial cluster center must be determined first to cluster by the FCM clustering algorithm. The former may be obtained by manually determining or defining an interval range and giving preference to the best cluster number. From the perspective of consumer clustering, the number of clusters is that of the evaluation channels, which is expressed as \( P \), since clustering better coordinates the needs of consumers that prefer different service channels.

The initial cluster center will change with the objective function value of the optimal fuzzy classification. However, it is difficult to meet the difference requirements between classes. In general, scholars add a penalty function to the objective function of the existing model (1) or similar models to maximize the between-class distance \( (1/p(p-1))\sum_{q=1}^{p}\sum_{k=1}^{p}d(y_{iq}, y_{jk}) \). However, the following problems are encountered:

(i) The inner-class distance is a function of the distances between each data point and the cluster center and the power of membership, while the between-class distance is the average value of distance differences between cluster centers. Both vary widely in value and thus are beyond comparison. Their
proceed. 

(i) Step 1: as the clustering result is sensitive to the selection of the initial cluster centers, the distance between cluster centers should be increased as much as possible. The dominant point with the most neighbors in dataset \( Z \) is taken as the first cluster center, the data point farthest from the dominant point as the second, the data point with the largest product of the distance to the two cluster centers as the third, and so on, until \( P \) initial clustering centers are solved.

(ii) Step 2: calculate \( u_{ik} \) and \( y_k \), respectively, by equations (2) and (3). \( \beta \) can be given or estimated by a sample variance:

\[
(1/p) \sum_{j=1}^{p} ((1/n-1) \sum_{i=1}^{n} (x_{ij} - (1/n) \sum_{i=1}^{n} x_{ij})^2).
\]

(iii) Step 3: set the threshold \( \alpha \) of the inner-class distance and \( \beta \) of the between-class distance. Variance \( s = \sqrt{(1/p-1) \sum_{k=1}^{p} (y_k - Y_k)^2} \) is used to characterize between-class differences, where \( Y_k = (1/p) \sum_{i=1}^{p} y_k \).

(iv) Step 4: if \( J(U, V) \leq \alpha \) and \( s \geq \beta \), then the iteration is terminated, and the obtained \( u_{ik} \) and \( y_k \) are the most suitable membership function and cluster center, respectively.

(v) Step 5 (sample classification): work out the distance \( \phi(x_i, y_k) \) between data points \( x_i, j \in (1, \ldots, m) \) and each cluster center, and classify the minimum values.

5. Simulation Research

Given that a service resource targets a large number of consumers and may be allocated via five channels, a random sample survey was conducted on 100 consumers to seek their service evaluation data on each channel. The consumer group is clustered to pursue a more effective allocation of resources.

By the given steps, the possibility of clustering each data point (consumer) is preprocessed based on the evaluation data. Calculate the distance between two respective evaluation data points by formula (4), and \( \varepsilon = 0.2 \) and \( y = 0.70 \):

(i) Data close to each other are initially clustered by Steps 1 to 4 in Section 3 to obtain Set \( Y (Y = F_1 \oplus F_2 \oplus \ldots \oplus F_m \oplus X) \) composed of \( F_1, \ldots, F_m \) and \( X \), where in \( F_1 = \{x_1, x_2\} \) and \( F_2 = \{x_{25}, x_{17}\} \) … \( F_i \) are regarded as new data and each data in \( X \) as separate data. Thus, the initial evaluation set is simplified to Set \( Y \) with 72 data points.

(ii) By Step 5, the data in Set \( Y \) are processed for connected graph \( G \) without circles (see Figure 1) formed by elements in \( Y \), wherein isolated points and points not on the main connected graph are not drawn.

(iii) Process each node on Graph \( G \) by employing steps 6 and 7 to conclude dataset \( Z \) (including 41 data points) on the basis of an adaptive nearest-neighbor classification rule. \( z_j \) represents one or several data points in \( X \), whose partial relationship is shown in Table 1.

(iv) Since the evaluation data involve 5 channels, find 5 initial cluster centers by Step 1 in Section 4:

\[
y^{0}_1 = (0.50, 0.42, 0.66, 0.53, 0.56),
\]

\[
y^{0}_2 = (0.96, 0.98, 0.17, 0.87, 0.06),
\]

\[
y^{0}_3 = (0.13, 0.01, 0.11, 0.04, 0.70),
\]

\[
y^{0}_4 = (0.88, 0.06, 0.59, 0.92, 1.00),
\]

\[
y^{0}_5 = (0.03, 1.00, 0.28, 0.26, 0.44).
\]

(v) Given \( m = 2 \), calculate \( u_{ik} \) and \( y_k \) with Steps 2 to 4. Suppose \( \alpha = 130.5 \) and \( \beta = 0.3 \). The iteration stops, provided that \( J = 130.43, s = 0.3227 \), and \( t = 14 \), so the cluster center is in its optimal state after 14 iterations:

\[
y^{14}_1 = (0.45, 0.32, 0.48, 0.24, 0.53),
\]

\[
y^{14}_2 = (0.92, 0.97, 0.18, 0.85, 0.09),
\]

\[
y^{14}_3 = (0.33, 0.19, 0.09, 0.30, 0.48),
\]

\[
y^{14}_4 = (0.87, 0.06, 0.59, 0.92, 0.99),
\]

\[
y^{14}_5 = (0.27, 0.76, 0.15, 0.17, 0.27).
\]

(vi) Perform Step 5 to classify samples, procuring a clustering result of Set \( Z \).

(vii) Get Set \( X \) based on Table 2 and the corresponding relationship between Set \( Z \) and Set \( X \) in Table 1. This is the clustering result of the original evaluation data, as shown in Table 3.

The distance \( ||V^{\text{present}} - V^{\text{previous}}|| \) (expressed by \( d_{ij} \)) upon 33 iterations is shown in Table 4, according to \[9\]. Changes in \( d_{ij} \) are shown in Figure 2.

Figure 2 illustrates that the value of \( d_{ij} \) increases first, decreases next, and then increases and is free of monotone convergence. The clustering center may not be optimal when
Figure 1: Connected graph of elements in Set Y.

Table 1: Correspondence between Set Z and Set X.

<table>
<thead>
<tr>
<th>Z</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_1$</td>
<td>$x_{31}, x_{38}$</td>
</tr>
<tr>
<td>$z_2$</td>
<td>$x_{38}, x_{77}, x_{92}, x_{97}$</td>
</tr>
<tr>
<td>$z_3$</td>
<td>$x_{63}, x_{76}, x_{95}, x_{96}$</td>
</tr>
<tr>
<td>$z_4$</td>
<td>$x_{56}$</td>
</tr>
<tr>
<td>$z_5$</td>
<td>$x_{12}, x_{19}, x_{46}, x_{90}$</td>
</tr>
<tr>
<td>$z_{41}$</td>
<td>$x_{77}, x_{48}, x_{98}$</td>
</tr>
</tbody>
</table>

Table 2: Cluster results of Set Z.

| C_1 | $z_8, z_9, z_{10}, z_{13}, z_{14}, z_{16}, z_{17}, z_{20}, z_{21}, z_{23}, z_{24}, z_{27}, z_{28}, z_{31}, z_{41}$ |
| C_2 | $z_{10}, z_{11}, z_{22}, z_{40}$ |
| C_3 | $z_1, z_2, z_3, z_{12}, z_{26}, z_{30}, z_{33}, z_{34}, z_{35}, z_{37}$ |
| C_4 | $z_{21}, z_{29}, z_{32}, z_{36}$ |
| C_5 | $z_{4}, z_{5}, z_{6}, z_{7}, z_{15}, z_{18}, z_{19}, z_{25}, z_{38}, z_{39}$ |

Table 3: Cluster results of Set X.

| C_1 (38) | $x_2, x_4, x_8, x_{10}, x_{12}, x_{17}, x_{20}, x_{23}, x_{25}, x_{29}, x_{30}, x_{32}, x_{34}, x_{37}, x_{38}, x_{39}, x_{42}, x_{44}, x_{45}, x_{51}, x_{55}, x_{57}, x_{59}, x_{61}, x_{62}, x_{64}, x_{67}, x_{68}, x_{73}, x_{75}, x_{78}, x_{83}, x_{84}, x_{85}, x_{89}, x_{91}, x_{98}, x_{99}$ |
| C_2 (8)  | $x_7, x_{15}, x_{24}, x_{43}, x_{50}, x_{60}, x_{82}, x_{88}$ |
| C_3 (26) | $x_1, x_3, x_6, x_9, x_{21}, x_{35}, x_{48}, x_{49}, x_{52}, x_{53}, x_{59}, x_{63}, x_{65}, x_{70}, x_{72}, x_{76}, x_{79}, x_{81}, x_{82}, x_{93}, x_{94}, x_{95}, x_{96}, x_{97}, x_{100}$ |
| C_4 (10) | $x_{5}, x_{11}, x_{18}, x_{27}, x_{31}, x_{40}, x_{41}, x_{47}, x_{50}, x_{71}$ |
| C_5 (18) | $x_{12}, x_{13}, x_{14}, x_{19}, x_{22}, x_{26}, x_{28}, x_{33}, x_{36}, x_{46}, x_{48}, x_{69}, x_{74}, x_{80}, x_{86}, x_{87}, x_{90}$ |

Table 4: $d_{i,j}$ after 33 iterations.

<table>
<thead>
<tr>
<th>i, j</th>
<th>$d_{i,j}$</th>
<th>i, j</th>
<th>$d_{i,j}$</th>
<th>i, j</th>
<th>$d_{i,j}$</th>
<th>i, j</th>
<th>$d_{i,j}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2</td>
<td>0.0130</td>
<td>9, 10</td>
<td>0.0276</td>
<td>17, 18</td>
<td>0.0024</td>
<td>25, 26</td>
<td>0.0023</td>
</tr>
<tr>
<td>2, 3</td>
<td>0.0141</td>
<td>10, 11</td>
<td>0.0249</td>
<td>18, 19</td>
<td>0.0021</td>
<td>26, 27</td>
<td>0.0024</td>
</tr>
<tr>
<td>3, 4</td>
<td>0.0157</td>
<td>11, 12</td>
<td>0.0199</td>
<td>19, 20</td>
<td>0.0019</td>
<td>27, 28</td>
<td>0.0027</td>
</tr>
<tr>
<td>4, 5</td>
<td>0.0177</td>
<td>12, 13</td>
<td>0.0134</td>
<td>20, 21</td>
<td>0.0019</td>
<td>28, 29</td>
<td>0.0029</td>
</tr>
<tr>
<td>5, 6</td>
<td>0.0205</td>
<td>13, 14</td>
<td>0.0085</td>
<td>21, 22</td>
<td>0.0018</td>
<td>29, 30</td>
<td>0.0032</td>
</tr>
<tr>
<td>6, 7</td>
<td>0.0239</td>
<td>14, 15</td>
<td>0.0055</td>
<td>22, 23</td>
<td>0.0019</td>
<td>30, 31</td>
<td>0.0035</td>
</tr>
<tr>
<td>7, 8</td>
<td>0.0272</td>
<td>15, 16</td>
<td>0.0038</td>
<td>23, 24</td>
<td>0.0020</td>
<td>31, 32</td>
<td>0.0039</td>
</tr>
<tr>
<td>8, 9</td>
<td>0.0287</td>
<td>16, 17</td>
<td>0.0029</td>
<td>24, 25</td>
<td>0.0021</td>
<td>32, 33</td>
<td>0.0044</td>
</tr>
</tbody>
</table>
the iteration discontinues when \(\|V^{\text{present}} - V^{\text{previous}}\| \leq \varepsilon\), and a more suitable center satisfying the conditions may appear after \(n\) iterations with smaller values of \(d_{i,j}\).

In this paper, the iteration ceased when \(J(U,V) \leq \alpha\) and \(s \geq \beta\), where the additional condition \(s \geq \beta\) ensured an appropriate distance between different classes and made it easier to attain appropriate values for \(\alpha\) and \(\beta\).

6. Conclusion

Complex huge group clustering is the basis for the effective distribution of service resources and group coordination; nevertheless, traditional FCM and its improved versions are incapable of processing numerous data points to be clustered. In this case, the deletion of data points was studied in the current paper by using a graph-based clustering algorithm, adaptive clustering algorithm, and Gaussian kernel clustering algorithm. Meanwhile, a new Gaussian algorithm was proposed to present both the inner-class distance and between-class distance, which the objective function fails to do.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References


