Logistics Service Provider Segmentation Based on Improved FCM Clustering for Mixed Data

Taoying Li, Yan Chen, Jinsong Zhang
Transportation Management College, Dalian Maritime University, Dalian 116026, P.R. China
ytaoli@126.com

***Abstract*—More and more logistics service providers turn up and make it is difficult to choose correct, economy and efficient ones for us. In order to achieve customer segmentation properly, we proposed an improved FCM algorithm by modifying distance function of categorical data and that of mixed data in this paper, and proved its theoretical correctness. We use clustering to reduce the range of logistics service providers and simplify the complexity of practice. Finally, we apply the proposed algorithm into logistics service provider segmentation, and its results show that the improved FCM algorithm is efficient and correct.**

***Index Terms*—clustering algorithm; fuzzy *k*-means clustering; logistics service provider segmentation.**

I. Introduction

With the development and perfection of the Logistics domain, logistics turns up everywhere. More and more logistics service providers turn up in order to satisfy people's demands, whose characteristics and mode of operation are different from each other. How to choose strong and proper ones among all logistics service providers is critical for producers and customers, which could decide the cost and efficiency of transportation. Thus, we classify logistics service providers with similar characteristics into one cluster and minimize the difference among logistics service providers in the same cluster, and maximize the differences of logistics service providers in different clusters. Then we can adopt different strategies for different logistics service providers.

Some existing papers provide a number of methods for customer segmentation. For example, Lijuan Huang employed SOFM neural network to classify customers [1]. Wei Gao adopted fuzzy clustering ensemble for customer segmentation in view of the uncertain factors [2]. Kai Peng used K-means clustering algorithm to classify telecommunications SMS business customers [3]. Yu-Jie Wang introduced fuzzy equivalence relation for customer segmentation [4]. Inspired from experience of customer segmentation in other domains, we use clustering algorithms to classify logistics service providers in this paper.

Clustering is to classify data points into clusters and makes the similarity of data points in the same cluster maximized and the similarity of data points from different data points minimized [5]. Clustering plays an important role in data mining, and could be widely applied into pattern recognition, computer visualization, fuzzy control, etc [6].

Due to the uncertain factors and mixed numerical and categorical data of logistics service providers, the improved FCM algorithm was introduced into logistic service provider segmentation. We modify the distance measuring method for categorical data due to the problem that existing methods couldn't get effective distance for categorical data. The logistics service provider segmentation based on improved FCM clustering was applied in practice and results shown that it is effective and suitable.

II. Degree of Correlation and Distance Measuring Methods for Categorical Data

Zhexue Huang presented a cost function for measuring the efficiency of clustering for mixed data [7], and it can be shown in (1).

$$F_4 = \sum_{j=1}^{n} \vartheta(d_j, c_l) \tag{1}$$

Here,

$$\vartheta(d_j, c_l) = \sum_{i=1}^{m_r} (x_{ji} - c_{li})^2 + \mu \sum_{i'=m_r+1}^{m} \delta(x_{ji'}, c_{li'})^2 \tag{2}$$

Here, m_r stands for the number of numerical attributes. At same time, we suppose that numerical attribute i could start from 1 to m_r and categorical attributes i' could start from m_r+1. Parameter μ represents correlation coefficient, and its range is (0, 1]. The majority of people think that $\delta(p,q)=0$ while p=q and $\delta(p,q)=1$ while $p \neq q$.

The clustering algorithm proved by Zhexue Huang considered both numerical and categorical data, however it has two shortcomings, one of which is that sum of all weights of numerical attributes is 1, another is that the definition of distance for categorical attributes couldn't reflect practice. For example, the distance between big and small is equal to that between big and middle, and equal to distance between middle and small, which does not conform to reality.

Clustering algorithm in [8] could be used to classify mixed numerical and categorical data, but the optimization process of its cost function is too complex that it is not suitable for large dataset.

doi:10.4304/jcp.7.11.2629-2633

We think that the distance couldn't be expressed clearly by $\delta(p,q)=1$ while $p \neq q$. We introduce the relationship degree n into the distance.

Relationship degree (RD) was definite in this paper ass follow: Supposing x_j and x_t are two arbitrary data objects, including m_c categorical attributes, and the $RD(x_j, x_t)$ is number that x_j and x_t are same in same attributes.

From the definition of relationship degree, we know that the maximum value of relationship degree is the number of categorical attributes. Therefore, we will present the theorem and prove it.

***Theorem*1**: Supposing x_j and x_t are two arbitrary data objects, including m_c categorical attributes, if $RD(x_j, x_t)=0$, supposing any object x_s, satisfying $RD(x_j, x_s)=a$ and $RD(x_t, x_s)=b$ $(a,b>0)$, then $RD(x_j, x_t)$ could be modified as $RD(x_j, x_t)=\min(a,b)/2$.

The proof for Theorem1 will be given after that for Theorem2.

***Theorem*2**: Let x_j and x_t be any two objects, including m_c categorical attributes, and the distance $\delta(x_j,x_t)$ between x_j and x_t could be obtain as (3).

$$\delta(x_j,x_t)=m_c-RD(x_j,x_t) \quad (3)$$

Proof: δ could be used to definite distance because it satisfies the characteristics of distance space.

1) Reflexivity: $\delta \geq 0$, and $\delta(x_j,x_t)=0 \Leftrightarrow x_j=x_t$.

2) Symmetry: $\delta(x_j,x_t)= \delta(x_t,x_j)$.

3) Transitivity: if $\delta(x_j,x_t) \leq \delta(x_t,x_s), \delta(x_t,x_s) \leq \delta(x_t,x_s) \Rightarrow \delta(x_j,x_t) \leq \delta(x_t,x_s)$.

4) Triangle theorems: $\delta(x_j,x_t) \leq \delta(x_j,x_s)+ \delta(x_s,x_t)$.

It is obvious that the reflexivity, symmetry and transitivity are established, and we only need to prove the triangle theorems.

Proof: Let x_j, x_s and x_t be any three different objects, and supposing all of them have m_c categorical attributes, if $RD(x_j, x_t)=0$, $RD(x_j, x_s)=a$ and $RD(x_t, x_s)=b$, then $RD(x_j, x_t)= \min(a,b)/2$, and there distance can be updated as follows.

$\delta(x_j,x_t)=m_c-RD(x_j,x_t)= m_c-\min(a,b)/2< m_c$

$\delta(x_t,x_s)+ \delta(x_t,x_s)=(m_c-RD(x_t,x_s))+ (m_c-RD(x_t,x_s))$

$= (m_c-a)+ (m_c-b)= m_c+ (m_c-a-b)$

Due to $RD(x_j, x_t)=0$, $RD(x_j, x_s)=a$ and $RD(x_t, x_s)=b$, then we can get $(m_c-a-b) \geq 0$, thus $\delta(x_t,x_s)+ \delta(x_t,x_s) \geq m_c$.

According to the proof mentioned above, we know that it is a distance space and could be used to obtain the distance between any two categorical data.

Proof for Theorem1: Let $RD(x_j, x_t)=x$, then we can know that $(m_c-a)+(m_c-b)>(m_c-x)$ due to triangle theorems, which can be shown in Figure 1. Next we can grasp $x>a+b-m_c$, at the same time to meet $|(m_c-a)-(m_c-b)|>(m_c-x)$. supposing $b>a$, we gain $x<m_c-(b-a)$.

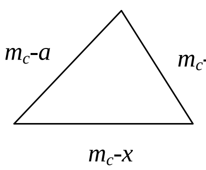

Figure1. Triangle theorems

Form the definition of relationship degree, we can establish the structure of relationship degree based on $RD(x_j, x_t)=0$, $RD(x_j, x_s)=a$ and $RD(x_t, x_s)=b$, which can be shown as Figure2.

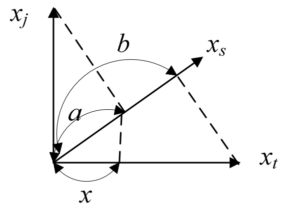

Figure2 structure of relationship degree among objects

Then we know that relationship degree satisfies the characteristics of geometric structure Form Figure2.

$b/m_c=x/a$, could be transformed to $m_c=ab/x$, substitution into $x>a+b-m_c$, we can get $x>a+b-ab/x$, then know that $x<a$ *according to* $x<m_c-(b-a)$.

Thus, in this paper, we let $x=a/2$. So $RD(x_j, x_t)= \min(a,b)/2$ meets the requirements.

We will give an example for relationship degree in Table1.

Table1. Categorical data

Object	*a1*	*a2*	*a3*	*a4*
x_j	X	A	T	M
x_s	X	A	S	N
x_t	Y	B	S	N

From Table1, we know that $RD(x_j, x_s)=2$ and $RD(x_t, x_s)=2$, and then we can get the distance $RD(x_j, x_t)=0$ by traditional function and $RD(x_j, x_t)=2/2=1$ by relationship degree mentioned above.

III. Improved FCM Clustering for Mixed Data Fuzzy K-Means Incremental Clustering Based on K-Center and Vector Quantization

A. Objective Function of Improved FCM Clustering for Mixed Data

In order to classify mixed data, the new cost function can be updated as (4).

$$F(T,W,C)=\sum_{l=1}^{k}\left[\frac{\sum_{j=1}^{n}\tau_{lj}\left(\sum_{i=1}^{m_r}w_{li}(c_{li}-x_{ji})^2+\mu\delta(c_l,x_j)\right)}{\sum_{i=1}^{m_r}(c_{li}-\bar{x}_i)^2+\delta(c_l,\bar{x})}\right] + \sum_{l=1}^{k}\gamma\left[\sum_{i=1}^{m_r}w_{li}\log w_{li}\right] \quad (4)$$

Where, $\sum_{l=1}^{k}\tau_{lj}=1, 1\leq j\leq n, \tau_{lj}\in\{0,1\}$ and $\sum_{i=1}^{m_r}w_{li}=1, 0\leq w_{li}\leq 1, 1\leq l\leq k$.

Here, $\bar{x}$ is the mean of all object, and $\bar{x}_i$ the value of the *i*th attribute, and $\bar{x}_i=\frac{1}{n}\sum_{j=1}^{n}x_{ji}$. The improved algorithm in this paper is efficient where $l > 1$. But the value of $\sum_{i=1}^{m_r}(c_{li}-\bar{x}_i)^2+\delta(c_l,\bar{x})$ can be zero, which means that the denominator of cost function may be zero and makes the computation for cost function impossible. At same time, the value of denominator may change at any time and is liner to square sum of distance between each mean $\bar{x}_i$ and $\bar{x}$.

The proof for the improved FCM clustering for mixed clustering is similar to that for numerical data in [9], but they have some following differences.

A. Weight correction

Let T and C be fixed, and F will be minimum while weight equals the process in (5).

$$w_{lt} = \exp(\frac{-\psi_{lt}}{\gamma}) / \sum_{i=1}^{m}\exp(\frac{-\psi_{li}}{\gamma}) \tag{5}$$

Where, ψ_{lt} satisfies (6).

$$\psi_{lt} = [\sum_{j=1}^{n}\tau_{lj}(c_{lt}-x_{jt})^2]/[\sum_{i=1}^{m_r}(c_{li}-\bar{x}_i)^2+\delta(c_l,\bar{x})] \tag{6}$$

B. Degree of membership

Let W and C be fixed, and we know that the jth object will belong to the lth cluster if their distance is closest, which can be shown as (7).

$$\tau_{lj} = \begin{cases} 1, & if \sum_{i=1}^{m_r} w_{li}(c_{li}-x_{ji})^2+\mu\delta(c_l,x_j) \\ & \le \sum_{i=1}^{m_r} w_{zi}(c_{zi}-x_{ji})^2+\mu\delta(c_z,x_j) \\ 0, & otherwise \end{cases} \tag{7}$$

Where, τ_{lj}=1means that the jth object belong to the lth cluster, and τ_{lj}=0 means that the jth object don't belong to the lth cluster.

The another form for solving degree of membership can be shown as (8).

$$\tau_{lj} = \frac{1/d(c_l,x_j)}{\sum_{z=1}^{L}1/d(c_z,x_j)} \tag{8}$$

Where,

distance $d(c_l,x_j)=\sum_{i=1}^{m_r}w_{li}(c_{li}-x_{ji})^2+\mu\delta(c_l,x_j)$.

C. Center selection

Center selection for mixed data can be divided into two aspects, one is for centers of numerical attributes, named means, another is for centers of categorical attributes, which can be obtain by the value of attribute maximum number of occurrences.

Which is similar to (9) or (10).

$$c_{li} = (\sum_{j=1}^{n}\tau_{lj}x_{ji}) / \sum_{j=1}^{n}\tau_{lj} \tag{9}$$

or

$$c_{li'} = (\sum_{j=1}^{n}x_{ji'})/n \tag{10}$$

Where, $1 \le l \le k$ and $m_r+1 \le i' \le m$.

B. The Solving Steps of Improved FCM Clustering for Mixed Data

The Solving steps of improved FCM clustering for mixed data in this paper can be presented as following.

Input: parameters m, m_r, n, k, γ, μ and the largest number of iterations.

Output: degree of membership T.

Step1. Let initial weight $w_{li}=m_r^{-1}$, choose k objects as k centers stochastically.

Step2. Obtain T according to (7) or (8).

Step3. Obtain cost function $F(T,W,C)$ according to (4).

Step4. Refresh C according to (9) and (10).

Step5. Refresh W according to (5) and (6).

Step6. Repeat Step2-Step5 until value of cost function unchanged or the time of iteration reaches to certain number.

The complexity of the clustering in this paper is O(mnk), which is liner to the number of objects needing to classify. Thus, the improved FCM clustering for mixed data could be suitable for large dataset.

IV. LOGISTICS SERVICE PROVIDER SEGMENTATION BASED ON IMPROVED FCM CLUSTERING FOR MIXED DATA

We classify 30 logistics service providers through their revolving credit, financial capacity, and customer evaluation, evaluation of bank and discount, which can be shown as Table2.

Table2 Data of 30 logistics service providers

No	Revolving credit	Financial capacity	Customer evaluation	Evaluation of Bank	Discount
1	0	0.4	0	0	1
2	0.2	1	0.2	0	0.4
3	0	0	0	0.2	1
4	0.2	0	0.8	0.2	0.2
5	0	0.2	0.8	0	0
6	0.2	0.2	0	1	0
7	0	1	0	0.4	0.2
8	0.8	0	0.2	0	0
9	1	0	0	0	0
10	0	0	1	0	0
11	0	0	1	0.4	0
12	0	0.2	0.8	0	0.2
13	1	0	0	0	0.4
14	0	0.2	0	0.2	1
15	0	1	0	0.4	0
16	0	0	0.2	0	1
17	0.2	0.4	0.8	0	0.2
18	0.2	0.4	0	0.2	1
19	0	0.8	0.2	0.4	0.2
20	0	0	0.2	0	0.8
21	1	0	0	0.2	0.4
22	0	0.2	0	0	1
23	1	0	0	0.2	0
24	0.4	1	0	0.2	0

25	0.8	0.2	0	0	0.2
26	0	0	0.2	0.2	1
27	0	0.4	0	0.2	1
28	0.8	0	0	0	0.2
29	0	0.2	1	0	0
30	0	0.8	0.2	0.2	0

We applied the improved FCM clustering for mixed data into dataset in Table2 and its result can be shown as Table3.

Table3 Clustering results

No	*k*	*F*	*c*
1	2	3.27	{0.41, 0.11, 0.05, 0.09, 0.63; 0.09, 0.49, 0.49, 0.23, 0.1}
2	3	0.88	{0.02, 0.18, 0.07, 0.1, 0.98; 0.06, 0.14, 0.89, 0.09, 0.09; 0.51, 0.43, 0.06, 0.21, 0.14}
3	4	0.07	{0.06, 0.14, 0.89, 0.09, 0.09; 0.02, 0.18, 0.07, 0.11, 0.98; 0.11, 0.82, 0.09, 0.37, 0.11; 0.91, 0.03, 0.03, 0.06, 0.17}
4	5	0.27	{0, 0, 0.15, 0.1, 0.95; 0.04, 0.32, 0, 0.12, 1; 0.2, 1, 0.2, 0, 0.4; 0.1, 0.8, 0.07, 0.43, 0.07; 0.49, 0.09, 0.46, 0.07, 0.13}
5	6	-0.19	{0.02, 0.18, 0.07, 0.11, 0.98; 0.1, 0, 0.9, 0.3, 0.1; 0, 0.15, 0.9, 0, 0.05; 0.11, 0.83, 0.086, 0.37, 0.11; 0.2, 0.4, 0.8, 0, 0.2; 0.91, 0.03, 0.03, 0.06, 0.17}
6	7	-1.3	{0, 0.07, 0.1, 0.1, 0.97; 0, 0, 1, 0.2, 0; 0.08, 0.2, 0.84, 0.04, 0.12; 0.07, 0.4, 0, 0.13, 1; 0.2, 0.2, 0, 1, 0; 0.1, 0.93, 0.1, 0.27, 0.13; 0.91, 0.03, 0.03, 0.06, 0.17}
7	8	-1.29	{0, 0.1, 0, 0.2, 1; 0.06, 0.14, 0.89, 0.09, 0.09; 0.05, 0.3, 0, 0.15, 1; 0, 0.4, 0, 0, 1; 0.08, 0.92, 0.08, 0.32, 0.08; 0.2, 0.2, 0, 1, 0; 0.2, 1, 0.2, 0, 0.4; 0.91, 0.03, 0.03, 0.06, 0.17}
8	9	-1.24	{0, 0.1, 0, 0.2, 1; 0, 0, 0.2, 0.07, 0.93; 0.05, 0.35, 0, 0.1, 1; 0.03, 0.1, 0.9, 0.1, 0.07; 0.2, 0.93, 0.13, 0.13, 0.13; 0.05, 0.75, 0.05, 0.55, 0.1; 0.2, 0.4, 0.8, 0, 0.2; 0.9, 0.05, 0, 0.05, 0.3}

The tendency of objective function *F* of the proposed algorithm can be shown as Figure3.

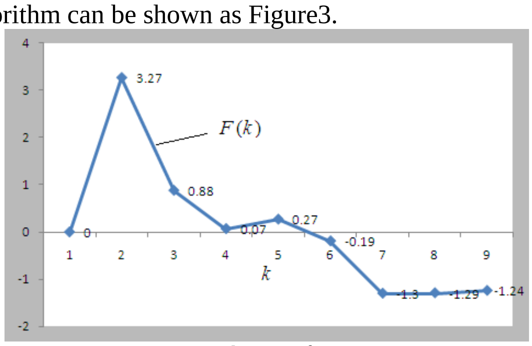

Figure3. Objective function *F*

We obtained 5 clusters as shown in Table4.

Table4 Results of logistics service provider segmentation

No	Number	No of provider	proportion
1	4	3, 15, 20, 26	13%
2	5	1, 14, 18, 22, 27	17%
3	1	2	3%
4	6	6, 7, 15, 19, 24, 30	20%
5	14	4, 5, 8, 9, 10, 11, 12, 13, 17, 21, 23, 25, 28, 29	47%

From Table4, we know that the 30 logistics service providers are divided into 5 clusters, which show that we can choose logistics service provider in certain group according to our preference. Clustering can reduce the range of logistics service providers and make us choose suitable logistics service provider from one group and improve our efficiency.

V. Conclusions

With the development of requirement for logistics in everyday life, all kinds of logistics service providers show up and how to choose suitable ones from so many logistics service providers is critical. We could improve the efficiency and applicability of logistics service provider segmentation if we can choose logistics service providers from a small range.

We study the improved FCM clustering algorithm by drawing lessons from customer segmentation in other fields and consider their characteristics in this paper. Then we update the distance function of categorical data. Finally, the proposed clustering is applied into logistics service provider segmentation and its results show that it is suitable for classifying logistics service providers and reducing their range.

Acknowledgment

This work was supported by the Fundamental Research Funds for the Central Universities (Grant No. 2011QN157) and the Major scientific research training project of Dalian Maritime University (Grant No. 2011ZD007).

References

[1] L. Huang, X. Gan. Customers Clustering Analysis and Corresponding Marketing Strategies based on SOFM Neural Network in e-Supply Chain. *System Engineering Theory & Practice*. 2007, 12: 49-55.

[2] W. Gao, C. He, X. Jiang. Customer Segmentation Study Based on Fuzzy Clustering Ensemble. *JOURNAL OF INTELLIGENCE*, 2011, 30(4): 125-128.

[3] K. Peng, Y. Qing, D. Xu. Customer Segmentation Modeling on Factor Analysis and K-MEANS Clustering. *Computer Science*, 2011, 38(5): 154-158.

[4] Y. Wang. A clustering method based on fuzzy equivalence relation for customer relationship management. *Expert Systems with Applications*, 2010, 37: 6421-6428.

[5] L. Jing, M. K. Ng, J. Z. Huang. An Entropy Weighting k-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data, *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 2007, 19(8): 1026-1041.

[6] T. Li, Y. Chen, J. Zhang, S. Qin. Incremental Clustering

Algorithm of Mixed Numerical and Categorical Data Based on Clustering Ensemble. *Control and Decision*, 2012, in press.
[7] Z. Huang. Clustering large datasets with mixed numeric and categorical values. Proceedings of *the First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, World Scientific, Singapore, 1997, 21-34.
[8] A. Ahmad, L. Dey. A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, 2007, 63: 503-527.
[9] T. Li, Y. Chen. A Weight Entropy k-means Algorithm for Clustering Dataset with Mixed Numeric and Categorical Data. Proceedings of *the Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, 2008, 36-41.

Taoying Li was born in Anhui Province, China on September 1983. Taoying Li received the BE degree in information management and information system at Dalian Maritime University, Dalian, China, in 2005, and recieved the PhD degree in Management Science and Engineering at Dalian Maritime University in 2010. She is currently a lecturer in the Transportation Management College, Dalian Maritime University.

She had carried out some projects and published several papers. She majors in Management science and Engineering and her research interests include data mining, system engineering, and artificial intelligence.

Yan Chen was born in Liaoning Province, China on December 1952. Yan Chen received the BE degree in computer software at Dalian Maritime College, Dalian, China, in 1978, the MS degree in computer application at Dalian Maritime College in 1989 and the PhD degree in Management Science and Engineering at Dalian University of Technology in 2000.

She is a professor in Transportation Management College, Dalian Maritime University, and the dean of the Key Laboratory in Liaoning Province on Logistics Shipping Management System Engineering. She has published three books and published more than 100 papers. Her main research interests include data mining, system engineering, special data mining, decision-making support system and data warehouse.