Research on the Fuzzy Clustering of Communication Data

Ting Gong[1, a], Hui Yan [1, b]

[1]*Department of Information Engineering Jilin Business and Technology College Changchun,China*

[a] *tingzi0505@sina.com,* [b]*yanhui7125@126.com,*

Abstract.

The ever-accelerated of communication technology narrows the distance among people. Telephone, as the main communication tool, has connected us quietly, and a huge social network is formed. The current call log data is used to summarize and analyze so as to make a reasonable decision, and thus to improve communication facilities, develop new communications businesses, and ensure that the communication company can keep its advantages to achieve more economic benefit.

Keywords: Call Data; FCM Clustering Algorithm; Fuzzy Clustering

Introduction

In recent years, communication business has developed prosperously, with the continuous intensification of market competition and the decrease of the telephone charge, the analysis on consumer behaviors becomes more and more important. The operating company should launch a new business for consumers according to their different needs, so as to achieve economic benefit. This paper divides the consumers based on their call logs, and the main difficulty is that they don't have a standard, which means that quantitative analysis cannot be taken and the specific number of categories is unknown. Therefore, based on the above features, a fuzzy

clustering analysis is adopted in this paper, and the consumers should be classified effectively through the validity control which decides the number of categories based on data features, further to deal with this kind of problem.

Fuzzy clustering analysis of communication data

In the field of information fusion, there are many definitions of information fusion. The quite exact concept is that: the multi-sensor information resources in different time and space are used, and computer technology is taken to analyze automatically, integrate, dominate and use the multi-sensor observation information by time sequence under a certain criterion, and to achieve the consistent interpretation and description of the measured object, so as to complete the decision and the estimated task, further to ensure that the system can achieve a more superior performance than its various components.

Establishment of the model

As for 300 consumers, we assume that the choices in 10 days are random, which means that call records in 10 days can represent their usual level, so the data we got is reliable.

As for the operators, the value of one user lies in the profit he or she brings to the company, and usually the profit made by users is related to their total call time(tm). Meanwhile considering that the existing telephone charge in market is mainly one-way charge system, so it is related to the proportion (rv) between calling and called, what's more, calling number c_d, called number and calls number per unit time of users are also the main factors for affecting the profits. Therefore, we choose four data as the characteristic variable of every user, according to the given parameter list, MATLAB is used to obtain these four data of every user, thus to work out the matrix X, X is the limited sample matrix under four-dimensional space. And then the FCM clustering algorithm is used to take the classification.

The principle of FCM clustering algorithm:

For the matrix X under four-dimensional space, we divide the users into c fuzzy classes, and put ci for the center of clustering in i class

$d_{ij} = \|c_i - x_i\|$ for the Euclidean distance between the center of clustering in i classes and the j-data point, the weighted index $m \in [0, \infty)$

Membership U for the matrix of $300 \times c$, and they meet the following formula:

$$\sum_{i=1}^{c} u_{ij} = 1 \quad \forall j = 1, ..., 300 \tag{1}$$

$$J(U, c_1, ..., c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{300} J_{ij}^{m} d_{ij}^{2} \tag{2}$$

FCM clustering algorithm is to get the solution when formula (2) taking the minimum in constrained conditions where formula (1) works. (1) represents the membership sums of every element belonging to their categories; formula (2) is the target function which judges the sum of distances, among which the number of categories c is given.

The steps of FCM algorithm are as follows:

Initialization: c for the number of cluster categories, $2 \le c \le n$ for the number of data, ε for the iteration stop threshold, m for the weighted index; the affiliated matrix U is initialized by random number whose value belongs to $[0,1]$, and it should meet the constraint condition of formula (1).

Step 1: calculate c clustering centers ci, i=1,..., c.

Step 2: calculate the objective function formula (2), when J is less than ε or relative to the last one, and the variation of J is less than ε, then stop.

Step 3: recalculate the affiliated matrix U, and return to Step 1.

The whole process is to make iterative changes on clustering center and classification matrix. The convergence of this algorithm has been proved. FCM algorithm can converge the local minimum point or saddle point of its objective function Jm (U, P) along an iterative sequence from any initial points.

Further improvement of the model:

Although users can be divided into several random categories through the above method, it will cause a problem: we don't know how many categories would be suitable and more effective. Therefore, in order to solve this problem, we introduce and adopt the fuzzy clustering validity function:

$$V_2^{PCM}(U;V;X)=\frac{\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^2\times d_{ij}^2}{n*(\min_{i\neq j}\{\|v_i-v_j\|^2\})}$$

Among which, n for the sample number, which is 300 in this subject, the subscript of $V_2^{PCM}(U;V;X)$ for weighted index 2 in FCM algorithm, dij for the distance between sample i and the j-clustering center. XIE-BENI can be explained as the ratio between the total square deviation of (U, V) and the separability index of V. The distance among various centers will be the largest with good effects of classification, which means that the separability index is larger. It seems that $V_2^{PCM}(U;V;X)$ will be the smallest when corresponding to the optimal class number n*.

According to the function, the steps of determining the optimal class number n*are as follows:

The range of c belongs to $[2,\sqrt{n}]$, which is given based on the using experiences of many researches and theory basis.

Calculate the value of V corresponding to every integer c when 2≤c≤" n.

Compare the value of every V, the value of c corresponding to the smallest V is the solution.

Model solving

The matrix manipulation is taken to judge the clustering validity, and FCM function is used to cluster above data, part of the code is as follows:

```
[center,U,obj_fcn] = fcm(B,c);
    maxU = max(U);
```

```
center
c
for i=1:c
C = find(U(i,:) == maxU)
end
```

The above results are from the calculation, and function c will be determined by cluster validity function.

Generally, set m=2, and the weights of the denominator is 1, when classnumber is [2,17], the validity index will get the following result, c=6, v=11.0643. And it can be sure that when c=6, V takes the minimum, that is the optimal class number n*=6. Thus there are 6 classes, and the clustering center matrix is:

$$\begin{bmatrix} 0.4 & 1651.9 & 8.9 & 7.8 & 9.3 \\ 0.5 & 1.279.5 & 7.0 & 6.1 & 6.0 \\ 6.4 & 357.5 & 2.0 & 2.1 & 1.6 \\ 3.2 & 605.5 & 3.3 & 2.6 & 2.2 \\ 1.4 & 865.1 & 4.7 & 3.6 & 3.4 \\ 0.2 & 2338.1 & 12.8 & 10.4 & 15.1 \end{bmatrix}$$

The matrix corresponds to the followings:

[Proportion between calling and called, total call time per day, calling number per day, called number, calls number]

The categories which users should belong to:

S_1 =[3 8 14 19 20 21 38 41 42 53 66 92]

S_2 =[4 11 12 16 18 24 27 34 47 49 51 52 54 61 70 71 75 87 93 103 137 144 170 172 190]

S_3 = [31 33 57 62 64 69 78 84 86 89 91 98 99 100 101 102 105 106 107 109 112 113 114 116 135 139 141 142 145 147 152 153 155 156 157 163 171 175 179 181 183 184 189 196 197 201 202 205 208 210 211 216 219 222 225 226 232 242 243 245 248 250 256 261 262 265 269 270 273 275 276 280 283 288 289 292 293 295 298 299 300]

S_4 =[10 22 25 28 36 39 45 48 50 60 63 65 68 72 74 76 77 81 82 85 88 90 94 95 96 97 108 110 111 115 121 122 123 124 125 126 128 129 130

131 133 134 136 140 143 150 151 159 164 166 169 173 177 180 182 185 187 188 191 192 193 194 195 198 199 204 212 214 218 220 221 223 224 228 233 234 235 237 238 240 241 246 251 252 253 254 257 258 263 264 266 267 268 271 274 278 279 281 282 284 285 286 290 291 294]

S_5 =[6 17 26 29 30 32 35 37 40 43 44 46 55 56 58 59 67 73 7980 83 104 117 118 119 120 127 132 138 146 148 149 154 158 160 161 162 165 167 168 174 176 178 186 200 203 206 207 209 213 215 217 227 229 230 231 236 239 244 247 249 255 259 260 272 277 287 296 297]

S6=[1 2 5 7 9 13 15 23]

Generalization of the model

The fuzzy cluster research of calling behavior analysis on communication users is a kind of new research field, and its method is a good method for data mining. And all base station of users can be classified through this method, and its rationality can be further judged by comparison. This method can be generalized to market classification, consumer classification, product classification and service classification, etc.

The calls data of this model roots from the communication company, and the company masters the calls behavior of consumers, and it ensures the coherence and continuity of their consumer behavior according to their use behavior and consumer tendency. Therefore, the analysis on classification standard and classification result provides communication company reasonable quantitative basis and advises. The communication company can make a few alterations on these results to obtain the call identify mechanisms.

Conclusion

The computer technology is taken to analyze automatically, integrate, dominate and use the multi-sensor observation information by time sequence under a certain criterion, and to achieve the consistent interpretation and description of the measured object, so as to complete the decision and the

estimated task, further to ensure that the system can achieve a more superior performance than its various components.

Acknowledgment

The authors gratefully acknowledge the funding of this study by The Jilin Province Department of Education research project(2013322) and the Jilin province talent development project in 2013 (20131079) and the Jilin province science and technology department soft science research project （20130420026FG）“Study of Jilin province engineering and technology talent support for strategy emerging industry”.

References

[1] HSUTH. An Application Of Fuzzy Clusteringin Group-positioning Analysis [J]. ProNatl Sci, Counc ROC(C), 2000, 10(2):157-167.

[2]Pan Yongli, The C_ Achievement of Fuzzy Clustering Analysis and Application, Yunnan: Journal of Yunnan Nationalities University [J], 2009(3):379-382.

[3]Zhou Xiaoguang, Analysis on Corporate Financial Risks Based on Fuzzy Clustering and Pattern Recognition, Science and TechnologyManagement Research [J], 2012(5):115-118.

[4]Zhang Li, Analysis on Long Staple Clustering Based on SPSS, Guangxi Textile Science & Technology[J], 2012(8):36-37

[5]XIE X, BENI GA,Validity Measure for FuzzyClustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(8):841-847.