Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Technology 6 (2012) 67 – 73

Procedia Technology

2nd International Conference on Communication, Computing & Security [ICCCS-2012]

Application of fuzzy clustering on software quality using max-min method

Jaya Pal [a]* , Vandana Bhattacherjee [b]

[a,b] *Department of Computer Science &Eengineering,Birla Institue of Technology,Ranchi,India*

Abstract

Fuzzy clustering analysis is a decision approach, makes fuzzy equivalent relation to classify objects into different clusters according to some criterion. In this paper it is applied in the clustering of software quality using the following steps: Firstly, fuzzy compatibility matrix is created using max-min method. Secondly, fuzzy equivalent relation is generated using composition relation. Finally, fuzzy equivalent matrix is used to obtain optimal threshold value to get clusters of software quality.

Kewords: Fuzzy Clustering analysis, software quality.

1. Introduction

Cluster analysis is a tool for data analysis and is branch in statistical multivariate analysis and also is an unsupervised learning technique in pattern recognition. Several methods of fuzzy clustering, such as fuzzy ISODATA [1], Fuzzy C-Means [2], Fuzzy K-nearest neighborhood Algorithm [3], potential based clustering [4] and others, have been proposed by various researches. The non-unique partitioning of the data in collection of clusters is the central idea in fuzzy clustering. The membership values of data points are assigned for each of

* Corresponding author. Tel.:+91-9931149772.
E-mail address: jayapal@bitmesra.ac.in (Jaya Pal), *vbhattacharya@bitmesra.ac.in* (Vandana Bhattacherjee).

doi:10.1016/j.protcy.2012.10.009

the clusters. The membership value of zero indicates that the data point is not a member of the cluster under consideration. Handling of extreme outliers in many crisp techniques are difficult but the tendency of fuzzy clustering algorithms is to give them very small membership value in surrounding clusters [5]. Thus fuzzy clustering provides a flexible and robust method for handling natural data with vagueness and uncertainty [6].Fuzzy clustering has been widely studied and applied in a variety of different areas[7,8,9]. In this paper, fuzzy clustering analysis is used for classification of software quality.

2. The principle of fuzzy clustering analysis

The fuzzy cluster analysis approach makes use of fuzzy equivalent relation to classify the objects into different criterion [10].The steps of fuzzy clustering is as follows:

Step 1: *To construct Fuzzy compatibility / primitive data matrix.*

Suppose universe U={x_1, x_2, ….., x_n} includes n objects to be clustered and each object has m factors to represent its characters. The basic principle of cluster analysis is to classified similar objects into same category. To obtain similarity coefficient r_{ij} between the objects x_i and x_j max-min method is used as

$$r_{ij} = \frac{\sum_{k=1}^{m} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{m} \max(x_{ik}, x_{jk})} \quad , \qquad (1)$$

Where $r_{ij} \in [0,1]$

Then fuzzy similarity relation matrix could be build up as

$$R = (r_{ij})_{nxn} \qquad (2)$$

The above matrix satisfies fuzzy compatibility relation. For each r_{ij} , if i =j then $r_{ij} = 1$, which indicates that R is called to have reflexivity. On the other hand, for each r_{ij} , $r_{ij} = r_{ji}$, which indicates that R is called to have symmetry.

Step 2: *To construct Fuzzy Equivalent Matrix*

A fuzzy relation which has symmetry, reflexivity and transitivity is called a matrix of equivalence relation. For clustering, fuzzy equivalence relation can be obtained by several composition computations.The procedure is : determine $R^2 = R \circ R$, determine $R^4 = R^2 \circ R^2$,……., determine $R^{2k} = R^k$ and stop, R^k is just fuzzy equivalence relation. i.e. the transitive closure t(R) of R equals to R^k

Step 3: To determine Optimal Threshold.

Threshold λ decides the separation between classes. Given $\lambda \in [0,1]$ and fuzzy relation $R = (r_{ij})_{nxn}$ then $R_\lambda = (\lambda r_{ij})_{nxn}$. λr_{ij} is called λ matrix(λr_{ij} doesn't mean multiplication of λ and r_{ij}).

$$\lambda\, r_{ij} = \begin{cases} 1 & , \quad r_{ij.} \geq \lambda \\ 0 & , \quad r_{ij} < \lambda \end{cases} \tag{3}$$

3. Application of fuzzy clustering analysis to software quality for category of optimal choice

Objects collection consists of 10 projects :U = {x_1,x_2 ,x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}}as shown in Table 1.In fuzzy cluster analysis, the project classified into the clusters should be the optimal choice.

3.1. Metrics used

Software quality factors may be enumerated as follows: portability, usability, reusability, correctness, maintainability etc. This paper focuses on quality of software using clustering and metrics were designed and / or adapted from Pal and Bhattacherjee [11] where the authors have developed a Fuzzy Logic System for prediction of software quality.

Description of metrics:

- *GUI (Graphical User Interface):* GUI was measured as the relative number of forms which were clearly displayed, on a scale of 0-10.
- *MEM (Meaningful Error Message):* MEM was measured as the relative number of meaningful error messages displayed by the software, on a scale of 0-1.
- *UM (User Manual):* UM was measured as the completeness of the user manual or help file, on a scale of 1-20.

The quality of the ultimate product (program) has been judged by team of three experts who ranked the various projects on a scale of 50-100 for usability and this served as the predicted output.

3.2 Data gathered

The ten projects are used to obtain the data. The statistics and a brief description related to each projects are depicted in Table 1.

1. The set up of the fuzzy compatible matrix

With formula (1) and data in Table 1 we can deduce the fuzzy compatible matrix R= (r_{ij}) $_{10x10,}$ i.e.

$$r_{ij} = \frac{\sum_{k=1}^{m} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{m} \max(x_{ik}, x_{jk})}$$

Sl No	Project Description	GUI	MEM	UM	SQ
1	Chatting on a Local area network	0	0.5	9	75
2	A tool for S/W Development Efforts Estimation	5	0.5	14	80
3	Asynchronous Linked File Downloader	1	0.4	8	72
4	Alpha to Itanium Migration of process control system	7	0.7	12	82
5	History Transaction Solutions	7	0.7	16	82
6	Online enterprise Resource Planning Based system (centralized inventory control solution)	6	0.6	14	83
7	Effective index based text retrieval system	7	0.8	18	91
8	Vulnerability Information Management System	1	0.2	9	62
9	Training and placement management for an institute	7	0.5	14	82
10	VPN for secure Information Exchange between LAN's	8	0.8	17	92

Table 1 Projects description and metrices, Graphics User Interface (GUI), meaningful Error Message (MEM), User Manual (UM), Software Quality (SQ)(ranks).

The steps of fuzzy clustering are as follows:

$$R = \begin{pmatrix} 1 & 0.48 & 0.80 & 0.48 & 0.40 & 0.46 & 0.36 & 0.87 & 0.44 & 0.36 \\ 0.48 & 1 & 0.48 & 0.80 & 0.82 & 0.94 & 0.75 & 0.52 & 0.90 & 0.75 \\ 0.80 & 0.48 & 1 & 0.47 & 0.39 & 0.45 & 0.36 & 0.88 & 0.43 & 0.36 \\ 0.48 & 0.80 & 0.47 & 1 & 0.83 & 0.85 & 0.76 & 0.51 & 0.89 & 0.76 \\ 0.40 & 0.82 & 0.39 & 0.83 & 1 & 0.86 & 0.91 & 0.43 & 0.90 & 0.91 \\ 0.46 & 0.94 & 0.45 & 0.85 & 0.86 & 1 & 0.79 & 0.49 & 0.94 & 0.79 \\ 0.36 & 0.75 & 0.36 & 0.76 & 0.91 & 0.79 & 1 & 0.39 & 0.83 & 0.92 \\ 0.87 & 0.52 & 0.88 & 0.51 & 0.43 & 0.49 & 0.39 & 1 & 0.47 & 0.39 \\ 0.44 & 0.90 & 0.43 & 0.89 & 0.90 & 0.94 & 0.83 & 0.47 & 1 & 0.83 \\ 0.36 & 0.75 & 0.36 & 0.76 & 0.91 & 0.70 & 0.92 & 0.39 & 0.83 & 1 \end{pmatrix}$$

2. The set up of the fuzzy equivalent matrix

Here we use composition computation to obtain transitive closure t(R) with square method as explained in Step 2 then we get t(R) is the fuzzy equivalent matrix as

$$R = \begin{pmatrix} 1 & 0.79 & 0.87 & 0.79 & 0.79 & 0.79 & 0.79 & 0.87 & 0.79 & 0.79 \\ 0.79 & 1 & 0.79 & 0.89 & 0.91 & 0.94 & 0.90 & 0.79 & 0.90 & 0.90 \\ 0.87 & 0.79 & 1 & 0.79 & 0.79 & 0.79 & 0.79 & 0.87 & 0.79 & 0.79 \\ 0.79 & 0.89 & 0.79 & 1 & 0.89 & 0.89 & 0.89 & 0.79 & 0.89 & 0.89 \\ 0.79 & 0.91 & 0.79 & 0.89 & 1 & 0.90 & 0.90 & 0.79 & 0.91 & 0.91 \\ 0.79 & 0.94 & 0.79 & 0.89 & 0.90 & 1 & 0.90 & 0.79 & 0.90 & 0.90 \\ 0.79 & 0.90 & 0.79 & 0.89 & 0.90 & 0.90 & 1 & 0.79 & 0.91 & 0.91 \\ 0.87 & 0.79 & 0.87 & 0.79 & 0.79 & 0.79 & 0.79 & 1 & 0.87 & 0.79 \\ 0.79 & 0.90 & 0.79 & 0.89 & 0.91 & 0.90 & 0.91 & 0.87 & 1 & 0.90 \\ 0.79 & 0.90 & 0.79 & 0.89 & 0.91 & 0.90 & 0.91 & 0.79 & 0.90 & 1 \end{pmatrix}$$

3. Clustering U with transitive closure of R

As for fuzzy equivalent matrix R, λ at different confidence level will be taken to carry out optimum clustering analysis. For each λ, the element in matrix $R\lambda$ will be substituted.

Firstly, let $\lambda = .91$ and we get

$$t(R)_{0.94} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

In $t(R)_{0.91}$, elements in any rows are not same to each other. Then elements of U are classified into ten clusters. Similarly intermediate matrices are calculated using different threshold values to get final matrices for obtaining optimal clusters as follows:

$$t(R)_{0.87} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}$$

Therefore, elements in U are classified into three categories, i.e.$\{x_{1,} x_3\}$, $\{x_2, x_4, x_5, x_6, x_7, x_9, x_{10}\}$ and $\{x_8\}$ It indicates that on lowering the cluster threshold value to obtain the optimal number of clusters according to the requirements.

4 Experimental results

K-Means algorithm, Fuzzy C-means algorithm and Max-Min method are applied to same data subset. The sum squared error (**SSE)** using Euclidean distance as

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{j=|c_i|} \left\| x_{j \in c_i} - \overline{x_i} \right\|$$

Where, $|c_i|$ is the cardinality of i th cluster, k is the number of clusters, $\overline{x_i}$ is centre of i th cluster.

The result is depicted in Table 2.

Table 2 Optimum result

	K-Means algorithm	Fuzzy C-Means Algorithm	Max-Min method
SSE	29.82	15.92	15.16

From this result, it is observed that the Max-Min method yields best clusters.

5. Conclusion and future research

In this paper Fuzzy clustering is applied to a set of Software Quality data and clusters are generated on using fuzzy threshold value in fuzzy equivalent matrix.. The advantage of fuzzy clustering is that even though we can allocate an absolute cluster membership to a data point, at a more fine granularity level and we can provide the percentage quality. As part of our ongoing work, we are collecting exhaustive set of data so as to

develop a model which can be for generalized use. Future research involves in collecting more data to serve as the basis of a generalized fuzzy tool for quality prediction.

Acknowledgements

The authors wish to thank the anonymous reviewers for their valuable suggestions concerning this paper. Support for this work from the postgraduate students of Birla Institute of Technology, and our colleague Mr. Partha Sarathi Bishnu is gratefully acknowledged.

References

[1] Dunn, J.C. "A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters", Journal of Cybernetics, 3, pp.32-57. 1973.

[2] Bezdek, J.C. "Fuzzy mathematics in pattern classification", Ph.D.thesis, Applied Mathematics Centre, Cornell University, Ithaca, 1973.

[3] Keller, J., Gray, M. R., and Givens, J.A. "A Fuzzy K-nearest neighbor algorithm", IEEE Trans. on Systems, Man and Cybernetics, SMC-15, 4, PP. 580-585, 1985.

[4] Chiu, S. L. "Fuzzy model identification based on cluster estimation", Journal of Intelligent Fuzzy Systems,2, pp.267-278, 1994.

[5] Looney, Carl G. "A Fuzzy clustering and Fuzzy Merging Algorithm", Available: http://citeseer.ist.psu.edu/399498.html.

[6] Thomas, B., Raju, G., and Wangmo, S "A modified Fuzzy C-Means Algorithm for Natural Data Exploration", World Academy of Science, Engineering and Technology 49 2009.

[7] M.S. Yang, A survey of fuzzy clustering, Mathematical and Computer Modeling, 1993, vol. 18, pp. 1-16.

[8] F. Hoppner, F. Klawonn, R. Kruse and T. Runkler, Fuzzy Cluster Analysis: Methods for Classification Data Analysis and Image Recognition, Wiley, New York(1999).

[9] V. Bhattacherjee and J. Pal , "Applying Fuzzy Clustering To Predict Software Usability", International Journal of Applied Research on Information Technology and Computing (IJARITAC), Volume:1, Issue:,Sept-Dec,2010.

[10] Qi Yang, " Application of Fuzzy Cluster Analysis to Tax Planning for Location of Foreign Direct Investment", IEEE 3rd international Conference on Information Management, Innovation Management and Industrial Engineering, PP. 553-555, 2010.

[11]Pal J., and Bhattacherjee, V., "A Fuzzy Logic System for Software Quality Estimation", in proceedings ICIT 2009, pp 183-187, 2009.