Construct Knowledge Structure Based on Clustering Algorithm

Jeng-Ming Yih

Abstract— **Fuzzy clustering algorithms are based on Euclidean distance function, which can only be used to detect spherical structural clusters. A Fuzzy C-Means algorithm based on Mahalanobis distance (FCM-M) was proposed to improve those limitations of GG and GK algorithms, but it is not stable enough when some of its covariance matrices are not equal. A new improved Fuzzy C-Means algorithm based on a Normalized Mahalanobis distance (FCM-NM) is proposed. Method integrates Fuzzy Logic Model of Perception (FLMP) and Interpretive Structural Modeling (ISM). The combined algorithm could analyze individualized concepts structure based on the comparisons with concept structure of expert. Integrated method and algorithm can help to analysis knowledge structure and cognition diagnosis. Cooperating fuzzy clustering and graphic structures analysis, that features of knowledge structures of each cluster are clearly displayed. Furthermore, Combined with fuzzy clustering algorithm based on normalized Mahalanobis distance could be very feasible for cognition diagnosis. Based on the findings and results, it shows that knowledge structures will be feasible for remedial instruction and this procedure will also useful for cognition diagnosis.**

Index Terms— **Fuzzy Logic Model of Perception (FLMP); Clustering algorithms; Interpretive Structural Modeling (ISM); Normalized Mahalanobis distance (FCM-NM)**

I. INTRODUCTION

The well-known ones, such as Bezdek's Fuzzy C-Means (FCM)[1], FCM algorithm was based on Euclidean distance function, which can only be used to detect spherical structural clusters. To overcome the drawback due to Euclidean distance, we could try to extend the distance measure to Mahalanobis distance (MD). However, Krishnapuram and Kim (1999) [2] pointed out that the Mahalanobis distance can not be used directly in clustering algorithm. Gustafson-Kessel (GK) clustering algorithm [3] and Gath-Geva (GG) clustering algorithm [4] were developed to detect non-spherical structural clusters. In GK-algorithm, a modified Mahalanobis distance with preserved volume was used. However, the added fuzzy covariance matrices in their distance measure were not directly derived from the objective function. In GG algorithm, the Gaussian distance can only be used for the data with multivariate normal distribution.We know Gustafson-Kessel clustering algorithm and Gath-Geva clustering algorithm, were developed to detect non-spherical structural clusters, but both of them based on semi-supervised Mahalanobis distance, these two algorithms fail to consider the relationships between cluster centers in the objective function, needing additional prior information. Added a regulating factor of covariance matrix, , to each class in objective function, the fuzzy covariance matrices in the Mahalanobis distance can be directly derived by minimizing the objective function, but the clustering results of this algorithm is still not stable enough. For improving the stability of the clustering results, we replace all of the covariance matrices with the same common covariance matrix in the objective function in the FCM-M algorithm, and then, an improve fuzzy clustering method, called the Fuzzy C-Means algorithm based on Normalized Mahalanobis distance (FCM-NM), is proposed. Zadeh developed fuzzy theory and it flourishes methodologies in many fields [5] [6]. One of these fields is cognition diagnosis and it help represent knowledge structure [7] [8] [9]. It is a common viewpoint that human knowledge is stored in the form of structural relationship among concepts and their subordinate relationship is fuzzy, not crisp. There are some methodologies for concept structure analysis but little is known about methodologies of individualized concept structure [10] [11] [12] [13]. Therefore, the development for methodology of individualized concept structure is an important issue and it is essential for cognition diagnosis and pedagogy [14]. In this study, the integrated method of individualized concept structure based on fuzzy logic model of perception (FLMP) and interpretive structural modeling (ISM) will be developed [15] [16] [17]. An example of empirical test data of linear algebra concept for students of learning deficiencies will also be analyzed and discussed. For the feasibility of remedial instruction based on the cognition diagnosis, clustering method is needed so that students within the same cluster own similar knowledge structures and students among different clusters have the most variance on knowledge structures [18] [19].

Manuscript received July 23, 2016

Jeng-Ming Yih, Min-Hwei Junior College of Health Care Management, Tai-Nan City736, Taiwan

II. LITERATURE REVIEW

A. Fuzzy Logic Model of Perception

Suppose there be a combination of two factors and . There are levels and levels for factor and respectively. The fuzzy true values are expressed as and . Fuzzy truth value and express the degree that the combination of and will support prototype [20] [21]. The probability that the combination of could be viewed as prototype can be expressed as follows [13] [22].

$$p(c_i\ ,\ o_j) = (c_i\, o_j)[c_i o_j + (1-c_i)(1-o_j)]^{-1} \qquad (1)$$

B. ISM Approach on Concept Structure Analysis

J. N. Warfield provided the foundation of ISM [23]. The integrated algorithms consist of three steps of algorithms, AMC, ASC and AFISM.

C. Fuzzy Clustering

Clustering technique plays an important role in data analysis and interpretation. Fuzzy clustering is a branch in clustering analysis and it is widely used in the pattern recognition field. Fuzzy clustering algorithms can only be used to detect the data classes with the same super spherical shapes. To overcome the drawback due to Euclidean distance, we could try to extend the distance measure to Mahalanobis distance (MD). However, Krishnapuram and Kim (1999)[24] pointed out that the Mahalanobis distance can not be used directly in clustering algorithm. Gustafson-Kessel (GK) clustering algorithm [25] and Gath-Geva (GG) clustering algorithm [26] were developed to detect non-spherical structural clusters. In GK-algorithm, a modified Mahalanobis distance with preserved volume was used. However, the added fuzzy covariance matrices in their distance measure were not directly derived from the objective function. In GG algorithm, the Gaussian distance can only be used for the data with multivariate normal distribution. To add a regulating factor of each covariance matrix to each class in the objective function, and deleted the constraint of the determinants of covariance matrices in the GK algorithm, the Fuzzy C-Means algorithm based on Mahalanobis distance (FCM-M),[27,28,29] was proposed, and then For improving the stability of the FCM-M clustering results, Replace all of the covariance matrices with the same common covariance matrix in the objective function in the FCM-M algorithm.

D. FCM-NM Algorithm

Not only z-score normalizing for each feature in the objective function in the FCM-CM algorithm, but also replacing the threshold D where

$$D = \sum_{i=1}^{c}\sum_{j=1}^{n}\left[\mu_{ij}^{(0)}\right]^{m}\left[\left(\underline{x}_j - \underline{a}_i^{(0)}\right)'\left(\underline{x}_j - \underline{a}_i^{(0)}\right)\right] > 0 \qquad (2)$$

With the determinant value of the crisp correlation matrix, and then, the new fuzzy clustering method, called the Fuzzy C-Means algorithm based on normalized Mahalanobis distance (FCM-NM) is proposed. We can obtain the objective function of FCM-NM as following:

$$J_{FCM-NM}^{m}(U, A, R, Z) = \sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m} d^2\left(z_j, \underline{a}_i\right) \qquad (3)$$

Ωηερε $X = [\underline{x}_1, \underline{x}_2, ..., \underline{x}_n]$, $\underline{x}_j \in R^p$, $j = 1, 2, ..., n$ (4)

$$\underline{z}_j = \left(z_{1j}, z_{2j}, ..., z_{pj}\right)', \; z_{ij} = \frac{x_{ij} - \overline{x}_i}{s_i}, \; j = 1, 2, ..., n, \; t = 1, 2, ..., p \qquad (5)$$

$$\overline{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}, \; s_i = \frac{1}{n}\sum_{j=1}^{n}\left(x_{ij} - \overline{x}_i\right)^2, \quad t = 1, 2, ..., p$$

(6)

Conditions for FCM-NM are

$$m \in [1, \infty); \; U = \left[\mu_{ij}\right]_{c\times n}; \; \mu_{ij} \in [0, 1], \; i = 1, 2, ..., c, \; j = 1, 2, ..., n$$

$$\sum_{i=1}^{c}\mu_{ij} = 1, \; j = 1, 2, ..., n, \; 0 < \sum_{j=1}^{n}\mu_{ij} < n, \; i = 1, 2, ..., c$$

(7)

$$d^2\left(\underline{z}_j, \underline{a}_i\right) = \begin{cases} \left(\underline{z}_j - \underline{a}_i\right)' R^{-1}\left(\underline{z}_j - \underline{a}_i\right) - \ln\left|\Sigma^{-1}\right| & if \left(\underline{z}_j - \underline{a}_i\right)' R^{-1}\left(\underline{z}_j - \underline{a}_i\right) - \ln\left|R^{-1}\right| \geq 0 \\ 0 & if \left(\underline{z}_j - \underline{a}_i\right)' R^{-1}\left(\underline{z}_j - \underline{a}_i\right) - \ln\left|R^{-1}\right| < 0 \end{cases}$$

(8)

The threshold of FCM-NM is a dynamic value rather than a constant, it is different from which of FCM-M in our previous work [33], and the convergent process is different from all of before mentioned algorithms.

E. Ordering Theory

Ordering theory (OT) is one of well-known ordering item algorithms for detecting ordering relationships between items, which was proposed by Airasian and Bart [16], Bart and Krus [18] in 1973. Ordering relationships and its precondition between items is obtained by constructing joint and marginal probabilities table.

III. EMPIRICAL ANALYSIS

The Mean clustering Accuracies of 100 different initial value sets of FCM□FCM-M and FCM-NM for the Dataset was shown in TABLE 2 and 3. From this table, we can find that the Accuracies of FCM is worse than the FCM-NM.

A. The Experiments with Iris Data

The balance Iris Data [30] with sample size 150 which features of the Iris data contains Length of Sepal, Width of Sepal, Length of Petal, and Width of Petal. The samples were assigned the original 3 clusters based on the clustering analysis. The characteristics of 3 clusters for Iris data were shown in Table 1.

TABLE I. The characteristics of 3 clusters for Iris Data

clusters	species
1	Setosa
2	Versicolor
3	Virginica

Each cluster has 50 sample points. The classification accuracies were shown in Table 2.

TABLE II. Classification accuracies of Iris Data

Algorithm	Accuracies (%)
FCM	89.33
FCM-M	90.00
FCM-NM	93.23

From Table 2, we find that the FCM-NM has the best result, up to 93.23%.

B. The Experiments with Abstract Algebra Datas

The test includes 19 items with 73 task-takers of university students. The data set comes from the National Taichung University of Education used in the empirical study with learning Abstract Algebra.

TABLE III. Classification accuracies of testing samples

Algorithm	Accuracies (%)
FCM	0.538
FCM-M	0.623
FCM-NM	0.654

The performances of our proposed FCM-NM algorithms, FCM-NM is simultaneously better than which of FCM algorithm in the datasets.

C. Knowledge Structures

There are 7 concept attributes within each item and they are depicted in Table 4. Although the combined algorithm of

FLMP and ISM could provide the concept structure of each task-taker respectively, it is unfeasible to display the concept structure of all task-takers in this paper.

TABLE IV. Concept Attributes of Test

Concepts	Concept Attributes
1	Groups and Subgroups
2	Normal Subgroups and Quotient Groups
3	Rings
4	Integral Domains
5	Ideals and Quotient Rings
6	Fields and Extension Fields
7	Galois Theory

Each item was selected in the AFISM step. The abstract algebra test is designed by the author for abstract algebra course of university students in this study. There are 73 third-year university students participating in the test. The algebra test includes 19 items and each item contains one concept. The concept attribute within each item are depicted in Table 5. All these items are dichotomous.

TABLE V. Item- Concept Matrix of Test

Item	Concept 1	Concept 2	Concept 3	Concept 4	Concept 5	Concep 6	Concept 7
1	1	0	0	0	0	0	0
2	1	0	0	0	0	0	0
3	1	0	0	0	0	0	0
4	1	0	0	0	0	0	0
5	0	1	0	0	0	0	0
6	0	0	1	0	0	0	0
7	0	0	1	0	0	0	0
8	0	0	1	0	0	0	0
9	0	0	0	1	0	0	0
10	0	0	0	1	0	0	0
11	0	0	0	1	0	0	0
12	0	0	0	1	0	0	0
13	0	0	0	0	1	0	0
14	0	0	0	0	1	0	0
15	0	0	0	0	0	1	0
16	0	0	0	0	0	1	0
17	0	0	0	0	0	0	1
18	0	0	0	0	0	0	1
19	0	0	0	0	0	0	1

After using FCM-NM clustering algorithm, we then classify the students into two groups, and then constructed ordering relationship knowledge structure among items for each group. From Fig. 1 to Fig. 2 are the knowledge structures for each group respectively. As show in Figure 1, the numbers on the right side are the correct ratios of the items of each level. With a comparison for these two knowledge structures, the group 1 has the most amounts of levels, and the group 2 has the least.

As show in Figure 1, the numbers on the right side are the correct ratios of the items of each level. The item 4 is on the bottom level of the knowledge structure. It means that, for some students, the item is more easier than others. Item 4 is belongs to concept attribute of Groups and Subgroups. Item 18 is belongs to concept attribute of Galois Theory.

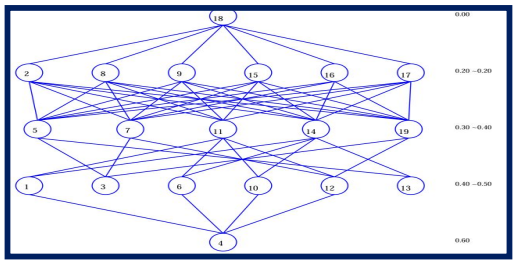

FIGURE I. Ordering Relationship for group 1.

As show in Figure 2, the numbers on the right side are the correct ratios of the items of each level. The item 10 and 14 are on the bottom level of the knowledge structure. It means that, for some students, the item is more easier than others. Item 10 is belongs to concept attribute of Integral Domains. Item 14 is belongs to concept attribute of Ideals and Quotient Rings.

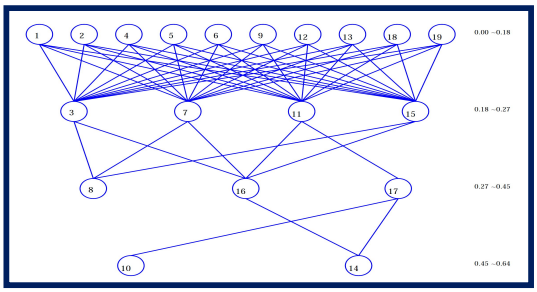

FIGURE II. Ordering Relationship for group 2.

As shown from Figure 3, one student is randomly selected from the sample. As to student 29, mastery of concept 1 which is 0.55 The concept of Groups and Subgroups is the basis for concept 2 ,3, 4, 5, 6, 7.

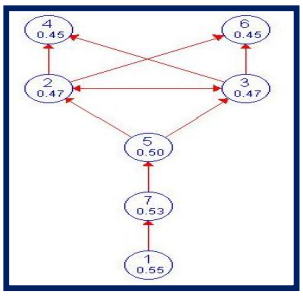

FIGURE III. KNOWLEDGE STRUCTURE OF STUDENT 29

As shown from Figure 4, As to student 71, mastery of concept 1 and 6 which is 0.50 The concept of Groups and Subgroups and Fields and Extension Fields are the basis for concept 2 ,3, 4, 5, 7

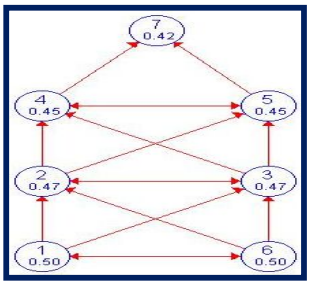

FIGURE IV. KNOWLEDGE STRUCTURE OF STUDENT 71

CONCLUSION

FCM is based on Euclidean distance function, which can only be used to detect spherical structural clusters. GK algorithm and GG algorithm were developed to detect non-spherical structural clusters. However, GK algorithm needs added constraint of fuzzy covariance matrix, GG algorithm can only be used for the data with multivariate Gaussian distribution. A Fuzzy C-Means algorithm based on Mahalanobis distance (FCM-M) was proposed to improve those limitations of above two algorithms, but it is not stable enough when some of its covariance matrices are not equal. An improved Fuzzy C-Means algorithm based on Normalized Mahalanobis distance (FCM-NM) is proposed. The experimental results of two real data sets consistently show that the performance of our proposed FCM-NM algorithm is better than those of the FCM algorithms.

Each cluster of data can easily describe features of knowledge structures. We can manage the knowledge structures of Abstract Algebra Concepts to construct the model of features in the pattern recognition completely. An integrated method of FLMP and ISM for analyzing individualized concept structure is provided. With this integrated algorithm, the graphs of concept structures will display the characteristics of knowledge structure. This result corresponds with foundation of cognition diagnosis in psychometrics [13]. This study investigates an integrated methodology to display knowledge structures based on fuzzy clustering with Mahalanobis Distances. In addition, empirical test data of abstract algebra for university students are discussed. It shows that knowledge structures will be feasible for remedial instruction [32]. This procedure will also useful for cognition diagnosis. To sum up, this integrated algorithm could improve the assessment methodology of cognition diagnosis and manage the knowledge structures of Abstract Algebra Concepts easily.

REFERENCES

1. J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum press, (1981).pp. 65-70, N.Y.
2. R. Krishnapuram and J. Kim, A note on the Gustafson-Kessel and adaptive fuzzy clustering algorithm, IEEE Transactions on Fuzzy Systems (1999). vol. 7 no. 4 August.
3. D. E. Gustafson and W. C. Kessel, Proc. IEEE Conf. Decision Contr. San Diego, CA, 761 ,1979.
4. F. Höppner, F. Klawonn, R. Kruse, T. Runkler, Fuzzy Cluster Analysis(1999). John Wiley and Sons,.
5. G. J. Klir, and B. Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentice-Hall (1995).New York, NY,.
6. L. A. Zadeh, Fuzzy sets, Information and Control, 1965, Vol. 8, pp. 338-353.
7. M. Smithson, and J. Verkuilen, Fuzzy Set Theory: Applications in the Social Sciences, Sage Publications (2006). Thousand Oaks, CA.
8. R. Coppi, P. Giordani and P. D'Urso, Component Models for Fuzzy Data, Psychometrika (2006). Vol. 71, pp. 733-761.
9. T. Sato, The S-P Chart and The Caution Index, NEC Educational Information Bulletin 80-1, C&C Systems Research Laboratories (1980). Nippon Electic Co., Ltd,. Tokyo, Japan.
10. J. P. Doignon and J. C. Falmagne, Knowledge Space(1999). Springer-Verlag.
11. K. VanLehn, Journal of the Learning Sciences(1999). Vol.8, p.71.
12. R. W. Schvaneveldt, Pathfinder Associative Networks (1991). Ablex.
13. W. P. Jr. Fisher, Rasch Measurement Transactions (1995). Vol.9, p. 442.
14. R. J. Mislevy and N. Verhelst, Psychometrika (1990). Vol. 55, p.195.
15. J. N. Warfield, Societal Systems Planning(1976). Policy and Complexity, Wiley.
16. J. N. Warfield, Interpretive Structural Modeling (ISM). In S. A. Olsen (Eds.), Group Planning & Problem Solving Methods in Engineering (1982).pp.155-201, Wiley,.
17. L. A. Zadeh, Information and Control (1965). Vol. 8, p.338.
18. Y. H. Lin, M. W. Bart, and K. J. Huang, Generalized Polytomous Ordering Theory(2006). [manual and software], National Taichung University, Taiwan.
19. T. Sato, Introduction to S-P Curve Theory Analysis and Evaluation (1985). Tokyo, Meiji Tosho.
20. G. Klir and B. Yuan, Fuzzy Sets and Fuzzy Logic, Theory and Applications(1995). Prentice Hall.
21. D. W. Massaro and D. Friedman, Psychological Review(1990). Vol. 97, p.225 .
22. C. S. Crowther, W. H. Batchelder and X. Hu, Psychological Review (1995). Vol.102, p.396.
23. J. N. Warfield, Crossing Theory and Hierarchy Mapping(1977).Vol.7, p. 505.
24. R. Krishnapuram and J. Kim, A note on the Gustafson-Kessel and adaptive fuzzy clustering algorithm, IEEE Transactions on Fuzzy Systems(1999). Vol. 7, no. 4 August.
25. D. E. Gustafson and W. C. Kessel, Proc. IEEE Conf(1979). Decision Contr. San Diego, CA, 761.
26. F. Höppner, F. Klawonn, R. Kruse, T. Runkler, Fuzzy Cluster Analysis(1999).John Wiley and Sons.
27. Gath, and A. B. Geva, Unsupervised optimal fuzzy clustering, IEEE Trans. Pattern Anal. Machine Intell(1989). Vol.11, pp.773-781.
28. Hasanzadeh R. P. R., Moradi M. H. and Sadeghi S. H. H., Fuzzy clustering to the detection of defects from nondestructive testing, 3rd International Conference: Sciences of Electronic Technologies of Information and Telecommunication(2005). March 27-31, Tunisia.
29. J. C. Dunn, A fuzzy relative of the isodata process and its use in detecting compact, well-separated clusters J. Cybern(1973). Vol.3, vol.3, pp. 32-57.
30. R. A. Fisher, The use of multiple measurements in taxonomic problems. Annals of Eugenics. Annals of Eugenics(1936). Vol.7, pp.179-188,.
31. B. Balasko, J. Abonyi and B. Feil " Fuzzy Clustering and. Data Analysis Toolbox For Use with Matlab" From http://www.mathworks.com/matlabcentral/fileexchange/7473.
32. K. K. Tatsuoka, and M. M. Tatsuoka, Computerized Cognitive Diagnostic Adaptive Testing: Effect on Remedial Instruction as Empirical Validation, Journal of Educational Measurement(1997).Vol. 34, , pp. 3-20.
33 H.-C. Liu, B.-C. Jeng, J.-M. Yih, and Y.-K. Yu, Fuzzy C-means algorithm based on standard mahalanobis distances. Proceedings of the 2009 International Symposium on Information Processing (ISIP'09), ISBN 928-952-5726-02-2.(2009). pp.422-427.