ISSN:2231-0711

Using of Two Analyzing Methods Multidimensional Scaling and Hierarchical Cluster for Pattern Recognition via Data Mining

Ali A. Ibrahim
College of Science
AL-Nahrain –University

Fwzi M. ALnaima
College of Engineering
AL-Nahrain –University

Ammar D. Jasim
College of Information Engineering
AL-Nahrain –University

***Abstract* -- The present study aims at making comparison between two analyzing methods , multidimensional scaling and hierarchical clustering methods ,on the other hand, to imply data mining by using of these two analyzing methods to classify and discriminate twenty five samples of technician pieces (Prehistoric goblets) through the study of Engineering shapes and special figures for every sample separately (case by case) , which backs to the periods before B.C. and discovered in Malaysia .The results of two methods concord and resemble each other equally in their classifications of Data. A multidimensional Scaling method seems to be more precise and give more details in comparison with hierarchical cluster methods.**

Keywords: MDS, HCA, Pattern Recognition, Data Mining

I- INTRODUCTION

From the beginning and ancient centuries, human being concerned with traces and Antiques, and considered as one of the basic symbols that referred to cultures at that time and witness of its progress and development.

The Engineering shapes and figures are major principles and basic symbols that discriminate and classify any technician piece (Prehistoric goblets) from another's, and also can express the Engineering shape by distances that may present these technician pieces in the best way. The primary utility of statistics is that they aid in reducing data into more manageable pieces of information from which inferences or conclusions.

The present research aims to classify and discriminate many of technician pieces and Prehistoric goblets by studying the Engineering figures of every pieces separately i.e. case by case by using two methods of analyses: scaling multidimensional scaling and hierarchical clustering, of twenty five types models of different Prehistoric goblets that come back to period before B.C. which discovered in Malaysia . To achieve and perform this goal , this paper divided into three sections, The first includes giving a central pictures about theoretical frame for both methods of analyses ,Multidimensional scaling and hierarchical cluster analysis. The second section concerned of practical part and discussion the results for both methods of analyses. The final section contains the conclusions that were derived from the results of the present research.

II- DATA MINING (DM)

Data mining is predicted to be "one of the most revolutionary developments of the next decade," according to the online technology magazine ZDNET News. In fact, the Massachusetts Institute of Technology (MIT) Review pointed that choosing of data mining will be one of ten emerging technologies that will change the word, which is processed under Knowledge discovery field. (2)

Data Mining is the analysis of (often-large) observational data sets to find unsuspected relationships and summarize the data in novel ways that are both understandable and useful to the user. (3, 4)

DM is one of the important steps of KDD process.. The common algorithms in current data mining practice include the following.

- Clustering: maps a data item into one of several clusters, where cluster are natural grouping of data items based on similarity matrices.
- Association rules: describes association relationship among different attributes.
- Summarization: provides a compact description for a subset of data.
- Dependency modeling: describes significant dependencies among variables

III- PATTERN RECOGNITION TECHNIQUE (5,6)

Data mining is the process of identifying patterns and relationships in data that often are not obvious in large, complex data sets. As such, data mining involves pattern recognition and, by extension, pattern discovery. Pattern recognition is most often concerned with automatic classification of characters.

The pattern recognition process starts with the unknown pattern, and ends with a label for the pattern. From an information-processing perspective, pattern recognition can be viewed as a data simplification process that filters extraneous data from consideration and labels the remaining data according to classification scheme.

The major steps in the pattern recognition processes are:

- **Features Selection.** Given a pattern, the first step in pattern recognition is to select a set of features or attributes from the universe available features that will be used to classify the pattern.
- **Measurement.** The measurement phase of the pattern recognition, involves converting the original shape into a representation that can be easily manipulated programmatically, depending on the underlying technology used to perform the pattern matching operation.
- **Features Extraction.** Features extraction involves searching for features in the data that are defines as relevant to pattern matching during feature selection. Clustering techniques, in which similar data are

grouped together, often form the basis of feature extraction.

- **Classification.** In classification phase of pattern recognition, data are classified based on measurements of similarity with other patterns. These measurements of similarity are commonly based on either a statistical or a structural approach. In the statistical approach, exemplar patterns are represented by points in a multidimensional space that is partitioned into regions associated with a classification. In the structural approach, the structures of exemplar patterns are explicitly defined.
- **Labeling.** The pattern recognition process ends when a label is assigned to the data, based on membership in a class.

IV- PROCESS REPRESENTATION

It is the purpose of data miner to use the available tools to analyze data and provide a solution to a problem. The data mining process can be roughly separated into three activities:

First step – Pre-processing (Input data):

We have the following input data:

$$A = \{x_i : x_i \in X\}$$

Where:

X: the sample space, i.e. the 20 goblets.

x_i : one goblet (one observation), and i=1,…,N

By symbol X, is:

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_N \end{bmatrix}$$

Second step – Threshold (making decision) : here, we use a specific formulas to make the following decision:

$$\text{Classification Techniques} = \begin{cases} x_i = 1 \text{ if } x_i \in X \\ x_i = 0 \text{ if } x_i \notin X \end{cases}$$

Which means, if x_i belongs to a specific group then it will be classified in this group otherwise will be classified in another one.

Third step – explaining (output):

The goal is to satisfy – classify the following:

$$B = \{x_i : x_i \in X\}$$

as matrix symbol X where:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m_1} \\ x_{21} & x_{22} & \cdots & x_{2m_2} \\ \vdots & & & \\ x_{h1} & x_{h2} & \cdots & x_{hm_\ell} \end{bmatrix}$$

Where:

h : No. of class in the X matrix.

m_l : size of class (No. of observation in a given class).

V- MULTIDIMENSIONAL SCALING METHOD

Multidimensional scaling is one of multivariable methods, which concerns with analyzing sample consist of (n) of elements (visible) scales with (p) of variables. (7)

Multidimensional scaling methods is defined as a series of styles (ways) designed to form figure (map) and, this figure shows indicates the relationships between (n) of groups of elements , depending on timetable of distances between these elements and this figure may be of one dimension (if the all elements are on one straight line) or of two dimensions (if the all elements are on the same horizontal (level)) or of three dimensions (if the all elements presented in space) or in numbers in numbers of higher dimensions multidimensions (in the case that the Engineering form is impossible). (8)

The multidimensional scaling starts with matrix of distances (as an input) among (n) of elements (number of these elements are (n-1) n ½ pairs of points) is SIJ , which presents distance of dissimilarity between the element (i) and the element (j) .

The matrix of distance can be calculated as follows:

$$d_{ij} = \sqrt{\Sigma_{k=1}^{p} (x_{ik} - x_{jk})^2} \quad \cdots (1)$$

Where :

dij : distance between the element (i) and the element (j)

xik : value of variable xk to element (i) .

xjk :value of variable xk to element (j).

Before starting calculating distances, it should be transformed the original variables to standardized variables , so all the variables (p) equated according to importance , in specifying these distances and to get the range of less multidimensional (t) (t<p) for the groups of points (n) to calculate the matrix from distances (dij) which has the same rank of the matrix origin distance (matrix of input) Sij ,to form and make line figure for (n) of elements , it should follow (9)

Prepare primary figure for (n) elements , and numbers of dimensions (t) (i.e.) , the symbols (x1 , x2 , x3 , …….., xn) to be supposed for every element in the range or space of dimensions of number (t) .

Calculating the descent distance (dij) (which presents the distance between element (i) and element (j)) among visible to this shape (shape of dimensions t), by calculating descent of smaller dual for (dij) on origin data distance (Sij) ,as well as evaluated (dij) through value (d^ij) which called disparities that is defined as the distance of chosen data to be similar as possible to distance of chosen figure (dij) .

Scale of co-efficient between distance of (chosen) figure (dij) and disparities (d^ij) , by using an appropriate statistics called efforts form (1) which is as follows :

$$STRESS\ 1 = \{\Sigma (d_{ij} - \hat{d}_{ij})^2 / \Sigma d_{ij}^2\}^{1/2} \cdots (2)$$

This statistics is defined as Scaling of land , which is by it impressed on the space figure for points to get as can as possible to distance of data (Sij)

Make a comparison between the values of efforts scaling with smallest value, if the □□□□□ of efforts scaling larger or greater than the smallest value, which would be find a new figure.

The multidimensional scaling using methods of steepest descent to find new figure which includes basically calculating derivations of efforts function, to decide the attitudes, to get little increase (improve the present picture in comparison with the original picture).

Table (1): Prehistoric data goblets discovered before B.C

No.	X1	X2	X3	X4	X5	X6
1	8	7	14	23	21	13
2	9	5	19	24	14	14
3	12	6	20	24	23	19
4	8	11	16	16	18	17
5	7	10	16	16	20	19
6	9	6	17	24	20	12
7	10	6	16	22	19	12
8	7	7	15	25	22	12
9	5	6	11	17	15	11
10	4	7	11	14	13	11
11	12	5	18	25	20	12
12	8	9	15	23	21	13
13	6	5	12	19	15	12
14	10	7	17	26	22	13
15	9	7	15	26	22	14
16	10	5	17	20	19	14
17	7	9	15	15	16	15
18	10	9	16	20	21	19
19	10	7	16	26	20	12
20	14	6	18	27	20	17
21	9	6	17	27	20	13
22	3	4	7	10	9	9
23	2	2	5	7	8	8
24	2	2	4	8	9	9
25	12	5	18	27	19	12

As soon as getting on persuasive of efforts value (less of the limited small value) therefore decrease the number of dimension by (one), and repeat the process (from the step 2 to step 5) until getting the less of number of dimension with acceptance value of efforts.

VI- HIERARCHICAL CLUSTER METHOD (10, 11, 12)

Hierarchical cluster is one of way of multivariable, since it takes part of grouping and classifying Data into clusters and these clusters are very similar to each other considering similarity within cluster.

There are many Algorithms that are using in cluster analysis, one of them (Hierarchic Algorithm) which is used in this work that starts calculating descent distance for every elements with other elements and to form groups through process of Agglomerates, and to start this process all the elements were single in different groups, then the closest elements start concord and combined gradually until at the end all the elements will be in one group.

Table (2): The frequency of ideal distances by using S-Stress formula. (1)

Iteration	S-stress	Improvement
1	.03761	.00979
2	.02871	.00889
3	.02751	.00121
4	.02713	.00038

(1)

Frequency stops because of the improving in value S – Stress was less than 0.001.

VII- RESULTS AND DISCUSSION

The input data was listed in Table(1) consist of Scaling with centimeters for six variables of twenty five Prehistoric goblets discovered in Thailand and were made before B.C.

A. Multidimensional Scaling Method

Table (1) shows the data manipulated by using Statistical software (SPSS) , and after ensure and confirm of precise of data, therefore implying the multidimensional Scaling technique, and the results were as the follows :

The iteration of calculation of S-stress formula on the ideal distances among twenty-five samples (Prehistoric goblets Discovered in B.C.) listed in table 2, which stops at iteration 4 and the value of improving in S-tress was less than 001.

S-stress scaling defines as the rate of disparities that is uncounted from multidimensional scaling, and this scaling helps in specifying and deciding the appropriate number from dimensions in analyzing (form map of analyzing) The results in table (2) construct figure (1) that decide and specify ideal number of dimension to be used in analyzing, and the best and less number of dimensions which is determined by equation 2 , where there was reduction the number of dimensions from six into only two new dimensions (variables)

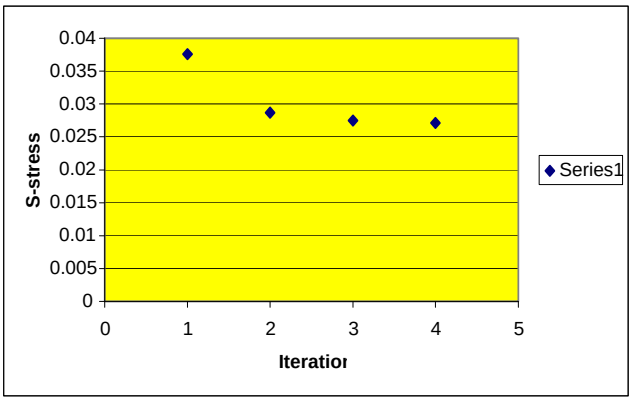

Figure (1) : The iteration using multidimensional Scaling Method where x-axis represents number of frequency of ideal distances by using S-Stress formula, meanwhile y-axis represents value of S-Stress.

Table (3) : Steps of clustering into different groups

Stage	Cluster Combined	
	Cluster 1	Cluster 2
1	23	24
2	11	25
3	19	21
4	1	12
5	14	15
6	6	7
7	9	13
8	14	19
9	1	8
10	4	5
11	6	16
12	22	23
13	1	14
14	4	17
15	9	10
16	1	6
17	3	20
18	1	11
19	4	18
20	1	2
21	1	3
22	4	9
23	1	4
24	1	22

Stimulus variables (samples of Prehistoric goblets) for twenty five samples according to first and second distances are construct the map that contains the location of twenty five prehistoric goblet, that represented the first layer, which is the output of Multidimensional Scaling Method.

B. Hierarchical Cluster Method

The data on table 1 manipulated by using Statistical software (SPSS) , and after ensure and confirm of precise of data, therefore implying the Cluster analysis technique, and the results were; the cluster combined (cluster 1 and cluster 2) which processed of agglomerate via Hierarchical Cluster analysis implying on twenty-five Prehistoric goblets, listed in table 3.

According to table 3, construct figure 2, which contains the steps of agglomerate via Hierarchical Cluster analysis implying on twenty five Prehistoric goblets, with each steps, it has a correspondence value that represent the distance (dissimilarity) between two clusters (observations), that represented the second layer, which is the output of Hierarchical Cluster Method.

C. Combine the Two Methods

Overlapping the two layers constructed figure 3 which represented the relationship between the two (multidimensional scaling and cluster analysis) methods, on

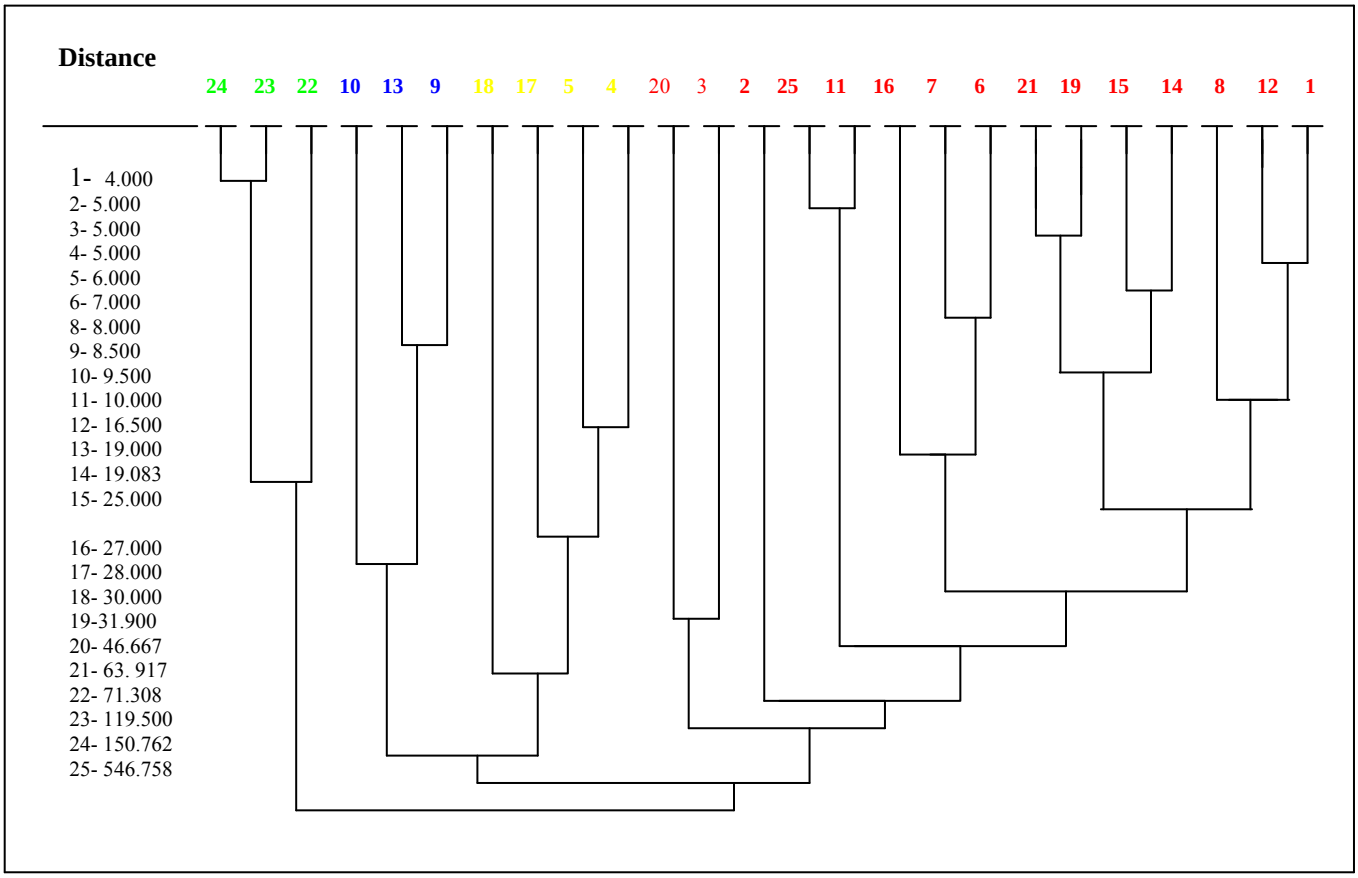

Figure (2): Steps of agglomerate via Hierarchical Cluster analysis implying on twenty five Prehistoric goblets, with each steps, it has a correspondence value that represent the distance (dissimilarity) between two clusters (observations) .Where read color represent group I, yellow color represent group II, blue color represent group III, and green color represent group IV.

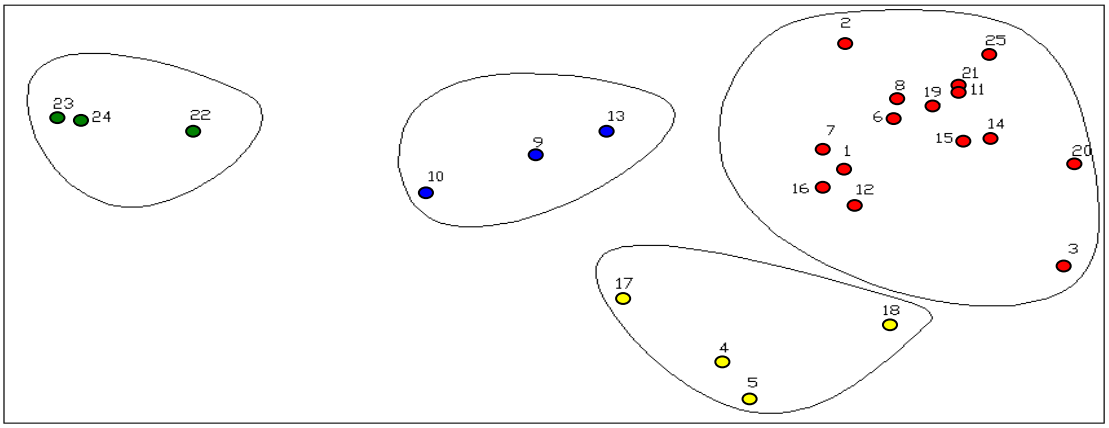

Figure (3): Constructed map of overlapping the two layers, which represented the relationship between the two (multidimensional scaling and cluster analysis) techniques.

the other hand this new map represents the classification and clustering of twenty five prehistoric goblet into four groups, as follows:
Contained, 1, 2, 3, 6, 7, 8, 11,12,14,15,16, 19,20, 21, 25.
Contained, 4, 5, 17, 18.
Contained, 9, 10, 13.
Contained, 22, 23, 24.

VIII- CONCLUSIONS

The multidimensional method proves to be more precise in giving more details than the Hierarchical cluster method, by specifying and classifying location of samples (Prehistoric goblets before B.C.) on map of two dimensions. While, The Hierarchical cluster method classifies different samples (Prehistoric goblets before B.C.) as different grouping without specifying location of these samples on the map.

The ability of two methods on specifying and classifying the different Engineering figures and scaling according to identical and similar scaling of different grouping

REFERENCES

[1] J. Natalia, C. Anastasova, "A Review of Multidimensional Scaling (MDS) and its Utility in Various Psychological Domains", Tutorials in Quantitative Methods for Psychology, University of Ottawa, Vol. 5(1), p. 1-10, 2009.
[2] T. Larose,"Data Mining Methods and Models", A John Wily and Sons, Inc Publication, 2009.
[3] A. George, "Application of Data mining in Medical Applications", thesis requirement for the degree of Master of Applied Science, Waterloo, Ontario, Canada, 2004.
[4] D.Hand,H. Mannila , and P. Smyth , "Principles of Data Mining" , The MIT Press., August 2001.
[5] B. Bergeron, "Bioinformatics Computing", Prentice Hall PTR, November 29, 2002.
[6] M.Subba Rao et al, "Comarative Analysis of Pattern Recognition Methods: An Overview", Indian Journal of Computer Science and Engineering (IJCSE), Vol. 2 No. 3 pp 385-390, Jun-Jul 2011.
[7] F. J. Manly, "Multivariate Statistical Methods APRIMER", 3rd Edition, Chapman and Hall CRC , 1986.
[8] N. Fieller, "Further Multivariate Analysis : Working Notes" , NRJF, University of Sheffield, 2001.
[9] M. Jansson, J. Johansson, "Interactive Visualization of Statistical Data using Multidimensional Scaling Techniques", Unilever and Institute of technology, 2003.
[10] J. Hair, B. Black, "Multivariate Data Analysis with Readings", 6th Edition, Prentice-Hall, Inc., 2008.
[11] M. Sollenborn, "Clustering and Case-Based Reasoning for user Stereotypes", thesis requirement for the degree of Master of Applied Science, Malardalen, University, Sweden, 2009.