Innovative Algorithm for Managing the Number of Clusters

Boumedyen Shannaq, Ibrahim Rashid Al Shamsi, Fouad Jameel Ibrahim AlAzzawi

Abstract*: This research work proposed an integrated approach using Fuzzy Clustering to discover the optimal number of clusters. The proposed technique is a great technological innovation clustering algorithm in marketing and could be used to determine the best group of customers, similar items and products. The new approach can independently determine the initial distribution of cluster centers. The task of finding the number of clusters is converted into the task of determining the size of the neural network, which later translated to identify the optimal groups of clusters. This approach has been tested using four business data set and shows outstanding results compared to traditional approaches. The proposed method is able to find without any significant error the expected exact number of clusters. Further, we believe that this work is a business value to increase market efficiency in finding out what group of clusters is more cost-effective.

Keywords: Clustering, Knowledge Management, Business Value, Segmentation Strategy.

I. INTRODUCTION

The Analysis of the client base remains incomplete considering that customers as a large bunch of identical people. Customers are divided into types and will use the product differently. Someone buys often, but not much, someone leaves quickly, someone buys a lot and often. To increase efficiency, you should find out what groups of clients are and then figure out how your actions will allow you to attract the clients you need. There is a common main way to understand your customer groups' so-called clustering. [1] Clustering combining similar objects into groups is one of the fundamental tasks in the field of data analysis and Data Mining. The list of application areas where it is applied is wide: image segmentation, marketing, anti-fraud, forecasting, text analysis, and many others. At the present stage, clustering is often the first step in analyzing data. After identifying similar groups, other methods are applied; a separate model is built for each group. The clustering problem in one form or another was formulated in such scientific areas as statistics, pattern recognition, optimization, machine learning. Clustering in Data Mining gains value when it is one of the stages of data analysis, building a complete analytical solution. It is often easier for an analyst to isolate groups of similar objects, study their features and build a separate model for each group than to create one common model on all data. This technique is constantly used in marketing, highlighting groups of customers, customers, products and developing a separate strategy for each of them. Very often, the data faced by Data Mining technology has the following important features: high dimension (thousands of fields) and large-volume (hundreds of thousands and millions of records) of database tables and data warehouses (extra-large databases); datasets contain a large number of numerical and categorical attributes. Most clustering algorithms involve comparing objects with each other based on some measure of proximity (similarity). A measure of proximity is a quantity that has a limit and increases with increasing proximity of objects. Similarity measures are "invented" according to special rules, and the choice of specific measures depends on the task, as well as on the measurement scale. As a measure of proximity for numerical attributes, the Euclidean distance calculated by the formula is very often used as described in [2]: D(x,y)=$\sqrt{Sumi(x-y)\text{^}2}$

To date, the number of methods for splitting groups of objects into clusters is quite large. However, we are interested in clustering algorithms in terms of their application in Data Mining. The model selection problem which represented as a clustering difficulty can be defined as the task of splitting a set of objects into a collection of equally disjoint subsets. Clustering is a separate problem which is considered as one of the most remarkable problems in unsupervised learning. Various methods have been proposed to solve the problem as suggested by [2]. From [3] Users of search engines, driving their requests into Google, Yahoo, and others, want to get an answer to their interest. In order to be satisfied and to find what the users were looking for, they need to put everything with regard to one need. Just like in a supermarket there are thematic goods: everything for a summer residence, barbecue, and holiday, etc. from a commercial point of view if the users find what they want will definitely come back. The same story with search queries, If we restrict ourselves to collect the semantics at the top and clustering it automatically using the principle "and so it will do" - the site visitor will not be satisfied. Put yourself in the user's place. Will you be loyal to the web resource if you spent time on it, "tearing through" a ton of non-targeted information and as a result did not find the answer to your question? Going to the site for various requests, people leave feedback. It is logical that if a person does not find the answer to his question on the page, he will not spend much time on it.

Revised Manuscript Received on January 15, 2020
***** Correspondence Author
Dr. Boumedyen Shannaq*, MIS, Business College, University of Burimi, Al Burimi, Oman. E-mail: boumedyen@uob.edu.om
Dr. Ibrahim Rashid Al Shamsi, Business College, University of Burimi, Al Burimi, Oman. E-mail: ibrahim.r@uob.edu.om
Dr. Fouad Jameel Ibrahim AlAzzawi, Iraq, Faculty Member-Al-Rafidain University College. E-mail: **fouad.alazzawi@ruc.edu.iq**

Most likely it will immediately close it and continue to analyze the issue further. The more common URL's between phrases, the stronger this connection. We have come to understand that analysis of the positions of search queries in search results allows us to obtain more accurate data for the distribution of search queries, depending on the goals and objectives. This research work proposes an adaptive clustering method (FANN-SOM) for a specific method, its efficiency is measured by the percentage of the data sets for which the number of clusters is correctly estimated, considering the incorrectly clustered instances and Time is taken to identify the number of the clusters. This approach has been tested on a set of critical clustering problems and shows excellent results compared to common approaches.

II. BACKGROUND AND LITERATURE REVIEW

[4] Currently, the amount of information in the public domain is growing at a tremendous speed. To simplify the navigation of these data, information retrieval systems, ranking methods, recommendation systems and much more are used. One of the approaches for organizing data is to build a catalog of topics, categories, subject index - these are all synonyms in this context. A similar solution found in each book there is a table of contents for the convenience of finding the necessary information. A similar hierarchical structure is often used to navigate sites. At present, there is an exponential increase in the need for information and at the same time there is a continuous replacement of information sources, the repositories of which often intersect. The search for the necessary information is becoming an increasingly complex, time-consuming and inefficient process. The user, moving from a list to a list of documents, the user should refine the search criteria and bring his query to some optimal set of words, by which he often receives a list of documents that is familiar to him in many ways. The search process goes in cycles, and the search time increases significantly. [5] Clustering (or cluster analysis) is the task of breaking down multiple objects into groups called clusters. Inside each group, there should be "similar" objects, and objects of different groups should be as different as possible. The main difference between clustering and classification is that the list of groups is not clearly defined and is determined during the operation of the algorithm. From the point of view of implementation, the system should use information from the user and from the search for documents. In particular, at the user's request, a list of documents found is built in a certain way and, using a request processed in a certain way. [6] The calculated matrix of document weights considering TF - IDF. Based on this information, it is necessary to obtain a list of headings and the distribution of documents over the retrieved documents, Fig.1, shows a sample clustering process.

	D1	D2	D3	D4	D5	D6	D7
D1	0	37.62978	14.8324	11.31371	31.60696	5.91608	8.602325
D2	37.62978	0	42.16634	35.69314	46.07602	36.7015	35.52464
D3	14.8324	42.16634	0	23.57965	20.27313	17.4069	21.07131
D4	11.31371	35.69314	23.57965	0	37.24245	7.28011	6.480741
D5	31.60696	46.07602	20.27313	37.24245	0	32.64966	34.21988
D6	5.91608	36.7015	17.4069	7.28011	32.64966	0	5.91608
D7	8.602325	35.52464	21.07131	6.480741	34.21988	5.91608	0

Fig.1. Sample of clustering process (Proximity matrix)

To solve this problem, it is proposed to perform fuzzy clustering of document vectors of the weight matrix. The problem is the dynamically changing number of components of these vectors depending on the modification of the request. It is also necessary to take into account the possibility of classification due to the possible emergence of new sources of information. There are a lot of methods for clustering documents, but there is no universal one. Each of them is good (or even optimal) under a number of conditions, but there are also disadvantages [2] is The Suffix Tree Clustering method [7] involves reprocessing the texts of documents. DS2L-SOM algorithm (Density-based Simultaneous Two-Level - SOM) presented in [8], the number of clusters obtained by various clustering methods (SSL = SOM+SingleLinkage, SKM = SOM+K-means). It is clear that the number of clusters is usually unknown factor that needs to be either stated by users based on their prior knowledge or expected in a certain way. A range of approaches has been proposed to evaluate the number of clusters. [9] Divided these approaches into two classes: global methods and local methods. The behavior of these methods has been compared in recent research [10]. [11] Performs a huge amount of computation and as a result of its work produces disjoint clusters. Examples of more recent developments in determining the number of clusters include an estimating approach using the approximated Bayes factor in model-based clustering developed by [12]. The concept of the Indexing method researched by [13]. [14] It is necessary to have a trained option and set the number of clusters. [15] In SingleLink, CompleteLink, GroupAverage the resulting clusters also do not intersect. The cluster borders are defined either by an "empty" region between clusters (i.e. large inter-cluster distances) or by a low-density region. The main parameter of K-means and Fuzzy-C-Means is to specify the number of clusters. For Kohonen SOM cards [16] the main drawback, as a rule, is the lengthy learning process. Although such methods that used the data segmentation obtained from the SOM is not optimal, since a part of information is lost during the first stage (dimensionality reduction) Various methods have been proposed to solve the problem Each of them is good (or even optimal) under a number of conditions, but there are also disadvantages.

III. FORMULATION OF THE PROBLEM

This application of cluster analysis, in general, is performed to the following basic steps demonstrated in fig. 2,

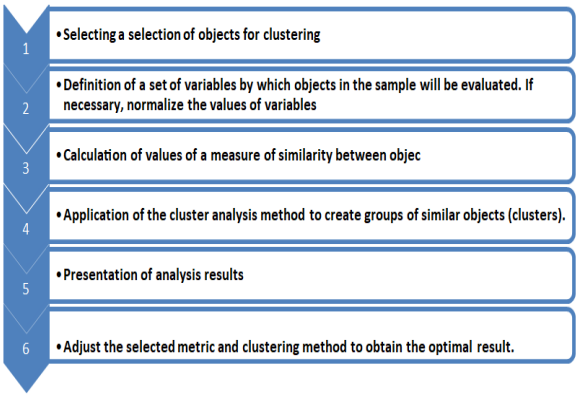


Fig. 2. Cluster analysis process

Let document **d** be represented by a set of keywords $\{kw_l\}^{M}_{l=1}$ that affect the classification of the document in any category, the document is the vector d = {kw1, kw2, ..., kw_M}. Then the power M of the term space KW will represent the dimension of the document space. The coordinates of the vector of the document that is submitted for processing will be the significance of a specific keyword for this document. If the keyword w_l is absent in some document, then the $\mathbf{l_{th}}$ coordinate of the vector is taken equal to 0.

In this research work, the measure tf * idf is taken as the basis, while the coordinates of the documents in the keyword space are converted into the matrix of the document weights. A cluster C in a keyword space KW is a set of similar objects, in the sense of a P measure, weight vectors of document keywords. Then the semantic similarity of the documents will be determined by the chosen P measure of similarity of the vectors representing them. Thus, it is necessary to cluster many documents such that $D=\{di\}^{N}_{i=1}$ .As a result, it is necessary to obtain a partition such that $C=C_1 \cup C_2 \cup C_3 \cup \ldots\ldots\ldots \cup C_k$ of the set D into possibly intersecting groups, where each document will be assigned the degree of belonging to each cluster, which determines the weight of the document's significance for each cluster of the heading. An important issue is the definition of the "quality" of clustering - the partition of many documents into the "correct" number of clusters. To date, there is no single approach to determine the best number of clusters. Only the upper and lower bounds on the number of K clusters are known. The boundaries of the interval, are usually taken as restrictions on the number of clusters such that $K \in [2, \frac{\sqrt{N}}{2}]$ where N is the number of documents.

IV. DESCRIPTION OF THE PROPOSED CLUSTERING METHOD

If To solve this problem in various sources, it is proposed to calculate the so-called objective function. The following are some of them and the conditions for finding the optimal number of clusters. Based on this formulation of the problem, to find the number of clusters and the corresponding partition, the following algorithm is formulated as described in fig.3,

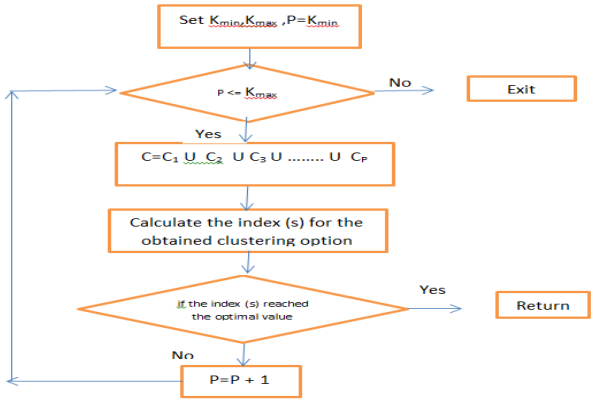


Fig.3 .clustering algorithm

Here, at each step of the algorithm, it is necessary to run some clustering method and calculate the values of the object functions. In this case, it is necessary to monitor the "jumps" of these object functions or to develop some algorithm for calculating the optimum of one or another object function, which is very problematic in this statement, since the object functions themselves contain information about the partition. It is very important to pay attention to the clustering options in which these "leaps" occurred. The next step is to look at the selected clustering options and select the best one. Obviously, the computational complexity of this approach increases sharply with increasing the input data. It should also be noted that when changing the volume of input data, it should restart this algorithm, which also increases the time complexity of the algorithm. [16] The classical Kessel algorithm, taking into account the forms of clusters, was used as a clustering algorithm. It is seen that, from the point of view of the optimality of the object functions, the best option is to split the original set into two clusters. As a confirmation of the correctness of the estimation of the number of clusters indicated above, we can note the fact that, when choosing a larger number of clusters, the values of the object functions do not fundamentally differ from the values obtained for the last variant of clustering.

A. Determining the Number of Clusters

In this work the self-organizing map SOM algorithm have been modified to determine the number of clusters as shown in fig.4,

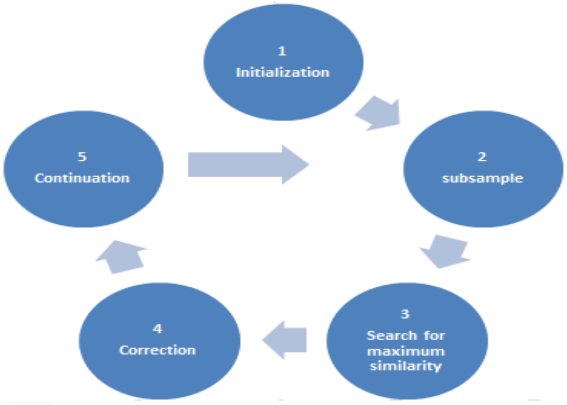


Fig. 4. Proposed strategy to identify number of clusters

1. Initialization:

Weight vectors wj (0), random values are selected. j = 1, ..., K, where K is the total number of neurons in the lattice.

2. The subsample:

Select a vector d with a certain probability, and let the dimension of the vector is M.

3. Search for maximum similarity:

The winner neuron id (x, y) is searched at step t using the minimum Euclidean distance criterion:

$$d(x,y) = \text{argmin}| d- y | , j= 1 , \ldots\ldots\ldots\ldots, K$$

Where (x, y) are the coordinates of the neuron in the lattice. At this step, it is necessary to take into account the problem of "dead" neurons. The following steps were used to solve it: Considering the fact that after the success the neuron "breaks", it is necessary to "limit" its activity at the next stage of training. To do this, one can keep track of the activity of neurons:

Pj (t +1) = {
Pj(t) + 1 / k , j <> id (x,y)
or
Pj(t) – Pmin , j = id (x,y)

Where p_{min} is the minimum "potential" allowing neuron to compete. In practice, p_{min} = 0.75 gives a good result. In addition, the number of neuron successes is taken into account when searching for the leader neuron, which allows us to use part of the neurons from a space region where there is no data or their number is negligible:

id(x,y) = argminj(NWj . | d – wj |) , j =1 ,……., K

Where NWj is the number of "successes" of neuron j.

4. Correction.

The vectors of semantic weights of all active neurons are corrected using the formula

Wj(t +1) = wj(t) + n(t)hj,id(x,y) (t)(d – wj (t))

n(t) : is the parameter of the similarity convergence rate

hj,id(x,y) (t) : function of the neighborhood function of the winning neuron id (x, y). Both of these parameters are dynamically changing:

- $n(t) = n_0 e^{\wedge} (-t/\tau_2)$
- $h_{j,id(x,y)}(t) = e \wedge (-d^2_{j,I} / 2\sigma^2 (t))$

Here n0 is the initial value of the convergence rate, the recommended value is 0.1, while n (t) should not be less than 0.01;

σ (t)= σ0e^((Q)) where Q=(-t/τl) and σ (t) is the width of the topological neighborhood of a neuron; at the initial stage, σ 0 is assumed to be equal to the radius of the lattice, which means the activity of all neurons of the network at the initial stage of training

τ2=1000 , τl= τ2/log σ0 - temporary parameters

5. Continuation. Return to step 2. Calculations continue until noticeable changes in the feature map stop.

The advantage of this algorithm is that it can independently determine the initial distribution of cluster centers. The adaptation process of the winning neuron allows one to adjust its weight in such a way that it becomes the "covariance center" of the found cluster. At this stage, the question remains about the size of the lattice of neurons. In this case, the task of finding the number of clusters is converted into the task of determining the size of the neural network. To solve it, the idea of hierarchical clustering methods was borrowed, where the partition begins to be built from a single cluster, which during clustering is split into a finite number of clusters in accordance with a certain condition. The applied Kohonen map sizing algorithm provides that for each cluster the value 1 is calculated,

$V_{ij} = 1/N_c \, SUMd_i \in C_{ij} |d_i – W_{ij}|$

Where C_{ij} is the set of vectors assigned to the K jth cluster, N_c is the number of documents in C_{ij}.

In this case, the entire learning process is necessary to determine the number of clusters, that is, a map change is possible. To determine its new sizes, we search for the neuron w *, for which vij is maximal:

$V_{max} = max(V_{ij})$

In order not to disturb the "training" of neurons, to the entire map is not rebuilt, only the row and column of neurons are added between the neuron w * and the neuron for which the distance, in the sense of the chosen metric, is the greatest.

Next, the weights of added neurons are initialized, for which the simplest method was used - calculating the weighted average weight of neurons from a topological neighborhood. The following criterion is proposed as a stopping criterion for resizing a map:

$N = 1/K_2 \; SUM_{i,j=1\ldots k} \; V_{ij}$

Where $N < 0.55 \; v_{pmax}$, v_{pmax} is the value determined at the previous step of specifying the map. For further research, it is necessary to include the semantic "isolation" of clusters in the condition for stopping the increase in network size. At the last stage, clusters are formed using the Gustafson Kessel algorithm .

B. Experiment

In order to show the advantages of the proposed work, we compare the proposed algorithm to three efficient algorithm available in weka Tool. And it's difficult to have an access to other developed algorithms. Weka tool offer three main clustering algorithms available as API interfaces as follow:

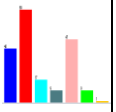
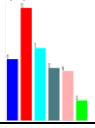
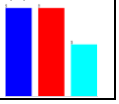
1-Cluster data using the capopy clustering algorithm found in [17].

2-Cobweb , found in [18] and [19].

3- Simple EM (expectation maximization, found in [20].

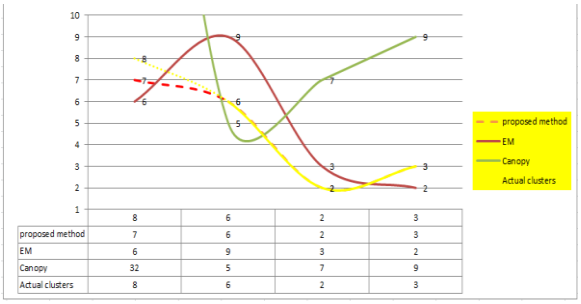
In this work we develop an API interface [21][22]23] to implement the proposed integrated approach by using Fuzzy Clustering algorithm and Artificial Neural Network Kohonen Self-Organizing Feature Maps (FANN-SOM). We use three factors to measure the performance of the proposed approach i.e. correct number of clusters, incorrectly clustered instances and Time taken to build model. The testing of the above algorithms was carried out on four test data, Table 1 describes the data set

Table- 1: data set information

Dataset Title	Instances& Sum of weights	Attributes	Class numbers & Distribution
CarShowroom.csv	194	30	(8) 
ElectronicMarket.csv	366	35	(6) 
FurnitureMarket.csv	768	9	(2) 
MobileMarket.csv	132	6	(3) 

After running the experiment the Cobweb clustering algorithm fail to identify the clustering numbers for the data set and cause memory problems during the testing model process .

Thus we remove this algorithm from the final results. Fig. 5, show the result correct number of clusters.



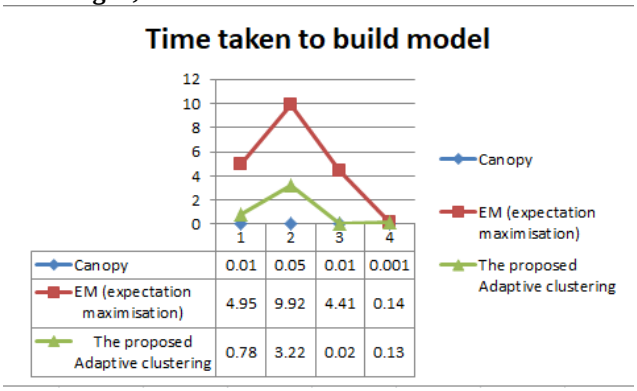
proposed method	8	6	2	3
EM	7	6	2	3
Canopy	6	9	3	2
Actual clusters	32	5	7	9
	8	6	2	3

Fig. 5. Results of correct number of clusters

Fig. 6, shows incorrectly clustered instances

Fig. 6. Results of incorrectly clustered instances

Fig. 7, shows Time taken to build the model.



	1	2	3	4
Canopy	0.01	0.05	0.01	0.001
EM (expectation maximisation)	4.95	9.92	4.41	0.14
The proposed Adaptive clustering	0.78	3.22	0.02	0.13

Fig. 7. Results of Time taken to build the model

V. CONCLUSION

This work generated data sets from different models that contain clear cluster structures with a known number of clusters. Four data set and three different methods capopy, Cobweb, and EM were tested and compared to the proposed work FANN-SOM. For a specific method, its efficiency is measured by the percentage of the data sets for which the number of clusters is correctly estimated, considering the incorrectly clustered instances and the time taken to identify the number of clusters. To examine the performances of these methods in the Weka tool application, we applied them to four real data sets and found that our proposed methods can correctly determine the best number of clusters. The outcomes for all the datasets show that the proposed adaptive method is able to find without any significant error the expected data division and the right number of what profitable groups of clients are. There is no doubt that such accurate and rich information capable of figure out serious market actions that allow business to attract the profitable clients they need.

REFERENCES

1. Boumedyen ,John ,Jambak .(2013) 'Taxonomic Knowledge Management Strategy for Managing Market Basketing , Journal of Current Computer Science and Technology ,Vol 3.
2. Jain, K. & Dubes, C. (1988)'Algorithms for clustering data'. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.
3. Douglas C. (2018)' Google Data Collection', Vanderbilt University August 15.
4. Boumedyen**,** S.'Adapt Clustering Methods for Arabic Document', American Journal of Information Systems, 2013, Vol. 1, No. 1, 26-30 Available online at http://pubs.sciepub.com/ajis/1/1/4 © Science and Education Publishing DOI:10.12691/ajis-1-1-4. http://pubs.sciepub.com/ajis/1/1/4/index.html
5. Alboukadel,K "Practical Guide To Cluster Analysis in R", sthda.com Edition 1 Unsupervised Machine Learning ,2017.
6. Boumedyen, S. ,Victor, A." Clustering the Arabic Documents (CAD)", Universal Journal of Applied computer Science and Technology (UNIASCIT),2011, Vol 1 (1), 05-08 .
7. Bezdek J. C." Cluster Validity with Fuzzy Sets" , Cybernet .1974, 3: 58–72.
8. Cabanes, G. , Younès ,B."Learning the number of clusters in Self Organizing Map", 2010, University of Paris 13, LIPN-CNRS France.
9. Sugar, C. A. and James. G. M. "Finding the number of clusters in a dataset: an information-theoretic approach". Journal of the American Statistical Association, 1998 ,750–763, 200
10. Tibshirani, R. ,Walther, G., Hastie, T. "Estimating the number of data clusters via the gap statistic". Journal of the Royal Statistical Society, 2001 , 63:411–423.
11. Wu, K. L. , Yang M.S. "A Cluster Validity Index for Fuzzy Clustering, Pattern Recognition Letters" , 2005 ,26: 1275–1291.
12. Fraley, C. , Raftery, A. E. "How many clusters? Which clustering method? Answers via model-based cluster analysis" ,Computer Journal, 1998,41:578–588.
13. Xie, X. L. , Beni, G. " A Validity Measure for Fuzzy Clustering", IEEE Trans. Pattern Anal. Mach. Intell, 1991,13 (8): 841–847.
14. Fraley, C. , Raftery, A. E. "Model-based clustering, discriminant analysis and density estimation". Journal of the American Statistical Association , 2002 ,611–631.
15. Campello, R. J, Hruschka, E. R." A Fuzzy Extension of the Silhouette Width Criterion for Cluster Analysis, Fuzzy Sets and Systems", 2006.
16. Valente, J. "Advances in Fuzzy Clustering and Its Applications", Wiley, 2007.
17. McCullum, A., Nigam, K., and Ungar, L.H. "Efficient clustering of high dimensional data sets with application to reference matching", In proceedings of the 6th ACM SIGKDD, 2000,167-178, Boston., MA .
18. Fisher, D. "Knowledge acquisition via incremental conceptual clustering",Machine Learning, 1987, 2(2):139-172.
19. Gennari, H. , Langley, P., Fisher, D."Models of incremental concept formation. Artificial Intelligence", 1990, 40:11-61.
20. Gupta and Chen, Y."Theory and Use of the EM Algorithm by Maya R", 2010.
21. Boumedyen Shannaq, AlShamsi,I.Saif,N,"Management Information System for Predicting Quantity Martials".TEM Journal, 8(4), 1143-1149, DOI: 10.18421/TEM84-06.(2019).
22. Al-Azzawi,F. & Boumedyen Shannaq. "Fuzzy Analysis Model for Classifying Exams Questions in Learning Quality Management System Based on Bloom's Taxonomy Verbs". 4. 69-79. 10.4206/aus.2019.n26.4.9, (2019).
23. Boumedyen Shannaq. "Innovative Web Service for Streaming Student Tweeting in Real-Time Technology", International Journal of Innovative Technology and Exploring Engineering, Volume-9 Issue-2, December 2019

AUTHORS PROFILE

Dr.Boumedyen Shannaq, is a skilled university professor for 10 years, primarily teaching Information Systems and Information Technology Courses. Interested in Electronic Learning, Innovation, Artificial Intelligence, Analysing Big Data. Professional skills in: Public Speaking, Time Management, Self-motivation, Inter-personal communication and Record keeping ,Focussed on Analyze ideas and use logic to determine their strengths and weaknesses.Manage oneself, people, time, and things.Obtained PhD in Information Systems (data Engineering-Technical Issues).

Dr. Ibrahim Rashid Al Shamsi, Ph.D. in Management, Currently Dr.Ibrahim is acting dean of Business college at university of Burimi.

Dr. Fouad Jameel Ibrahim, PhD Faculty Member -Al-Rafidain University College Quality Officer Baghdad-IRAQ mobile