Int. J. Nonlinear Anal. Appl. 14 (2023) 8, 55–66
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.28664.3950

Development of FCM method to increase clustering accuracy in big data

Hadis Changizi[a], Alireza Pour Ebrahimi[b,*], Mohammad Ali Afshar Kazemi[c], Reza Radfar[d]

[a]Department of Information and Communication Technology Management, Qeshm Branch, Islamic Azad University, Qeshm, Iran
[b]Department of Industrial Management, Karaj Branch, Islamic Azad University, Karaj, Iran
[c]Department of Industrial Management, Tehran Branch, Islamic Azad University, Tehran, Iran
[d]Department of Industrial Management, Science and Research Branch, Islamic Azad University, Tehran, Iran

(Communicated by Mohammad Bagher Ghaemi)

Abstract

Due to the spread of the Internet and its pervasiveness, "big data" is created daily. Processing this amount of data requires a system with high processing power. In fact, the production and collection of data from a wide range of different equipment and tools lead to the creation of large-scale databases. In dealing with large and unstructured databases and their management, there are always challenges. This study aims to present a model to increase the clustering accuracy of big data using a fuzzy clustering system based on data mining in a MatLab programming environment. For this purpose, first, the importance of each variable in the decision tree models in SPSSModeler software is determined, then with the help of these results, fuzzy rules are explained and a fuzzy inference system is formed in MATLAB software. This study uses data mining techniques such as C&R Tree, Chaid and C5.0 to study the development of the FCM method to increase clustering accuracy in high volume data and related factors such as data preparation indicators, data type Data quality, data dimensions, data volume and number of clusters were evaluated as inputs and clustering accuracy index was evaluated as output. Then, with the help of these results, the rules of forming a fuzzy inference system were determined and by explaining the membership functions of the decision model, it showed what effect each input index has on the output index.

Keywords: FCM, big data, clustering, fuzzy, data Mining
2020 MSC: 62R07

1 Introduction

Due to the spread of the Internet and its pervasiveness, a "large amount of data" is created daily. Processing this volume of data requires a system with high processing power [10]. Conventional systems today are not able to process such large volumes of data, and on the other hand, providing systems with high processing power due to high costs is beyond the reach of businesses [14].

*Corresponding author
Email addresses: uni.changizi@gmail.com (Hadis Changizi), support@apebrahimi.com (Alireza Pour Ebrahimi), dr.mafshar@gmail.com (Mohammad Ali Afshar Kazemi), radfar@gmail.com (Reza Radfar)

Received: October 2022 *Accepted:* December 2022

Extraction of repetitive patterns has consequences such as space complexity [7] and [4] and [11]: Input data, intermediate results and output patterns can be used for "memory placement". They are too large to prevent the execution of a large number of algorithms, as well as time complexity: Many existing methods rely on complete search or complex data structures to extract duplicate patterns, which is unworkable for large data sets [10].

In interacting with large and unstructured databases and its management, there are always many challenges [8] and [6], but the leading challenges are obstacles to work with large data, but do not extract value. It does not return from them [3]. On the other hand, understanding information about human behavior by exploring petabytes of network data suggests a tendency to research social behaviors and demonstrates the importance of Internet software design and service development [17] and [16].
At the same time, the implementation of mobile networks that produce huge data can be the best social sensor for these studies [12] and [5]. One way for web users to get to the subject more quickly and get a comprehensive view of it is to increase the accuracy of clustering in large data, one way is to use "clustering" [5] and [9].

Big data refers to a set of large and complex data that cannot be processed by traditional data processing software or is difficult [15]. The main challenge includes data analysis, collection and search [2]. The widespread use of multi-way sensor technology, as well as Big data, has made the need for tools for data analysis and analysis strongly felt [7] and [4]. On the other hand, matrices due to their limitations in data size and storage can not meet these needs well, so the use of tensors and tensor analyzes to collect and analyze large data is important (2019) [1]. The use of tensors makes it possible to move towards models that are essentially polynomial and their uniqueness, unlike matrices, is established under moderate and natural conditions [13].

Bu et al. Presented an efficient c-means method based on data analysis for clustering big data on the Internet of Things (IoT). The results show that the developed design significantly achieves clustering efficiency by increasing the clustering accuracy compared to the traditional algorithm, which indicates the potential of the developed design for exploring and extracting intelligent data from IoT big data. Yang et al. Examined the use of large-scale data analysis to cluster interest in products and services and link clusters to financial performance. In doing so, the analysis of existing data, such as email, which all companies have, can be used to validate new methods of identifying and monitoring product demand, informing marketing strategies, using Big data analysis. Zhang et al. Examined the c-means weighting algorithm on the cloud for big data clustering. The results show that the proposed scheme performs more efficiently than the traditional c-means weighted algorithm and achieves good scalability over the cloud for big data clustering.

In fact, the problems of this research can be the ambiguity and fatigue of decision-makers to increase the accuracy of clustering in big data (electronic businesses active in the field of banking and professors and students in the fields of industrial management, banking and IT) due to the combination of different methods of data services. IT Business, Big Data Clustering Business Services Design Stage, Big Data Clustering Business Services Transfer Stage, Big Data Clustering Business Services Operation Stage, and Big Data Clustering, Continuous Improvement of IT Services. On the other hand, the need to use fuzzy clustering system based on modelling data mining increases the accuracy of clustering in big data to increase trust and confidence in decision making, as well as the need for multiple expertise through the simultaneous application of multiple knowledge areas. There are problems with increasing clustering accuracy in big data. In the present study, a fuzzy clustering system based on data mining to increase the clustering accuracy of big data, called BD-Cluster FCM is presented for the first time in the field of related research. According to the main issue in this dissertation, the following main research question can be asked to better understand the research issue: How can the increase of clustering accuracy in big volume data be comprehensively modelled?

2 Method

The method of this dissertation is analytical-modelling in terms of purpose because, on the one hand, the concepts related to big data are described in detail and on the other hand, the relationships between these concepts are evaluated and determined by experts. Concepts and variables related to big data have been extracted from books and other library resources, and those variables and concepts have been evaluated using the opinions of experts. The decision tree is specified in SPSSModeler software, then with the help of these results, fuzzy rules are explained and a fuzzy inference system is formed in MATLAB software. The spatial scope of the present study is electronic businesses active in the field of banking in the country. Due to the use of articles and documents from different sources, the method of data collection in this dissertation is "case study of data" and to increase the accuracy of clustering that is extracted from bank data, bank data is used. Figure 1 shows the research steps:

Defining qualitative variables using language constraints and assigning fuzzy numbers and sets and membership functions to them.

Modelling domain concepts Increasing the clustering accuracy of big data to identify the importance of input and output variables and plot the relationships between them using decision tree models in SPSSModeler software.

Designing an inference/expert system using artificial intelligence calculations based on definitions and design using MatLab programming environment, which includes extracting expert rules and evaluating them by experts and creating a database of rules and de-fuzzing.

Analysis of output of fuzzy clustering system based on data mining to model the increase of clustering accuracy big data.

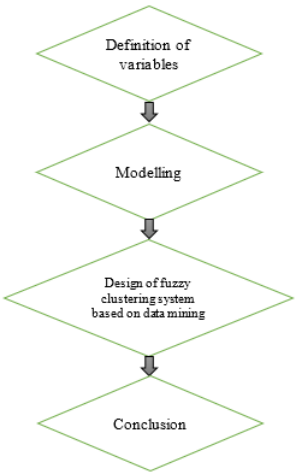

Figure 1: Flowchart of research step

2.1 Statistical society

Due to the combined methodology of fuzzy clustering system based on data mining in MatLab programming environment and IBM Modeler (Clementine) data-mining environment in modelling research, increasing the clustering accuracy in big data will be used by experts. The study population of this dissertation can be divided into two general groups: the first group includes academic professors in the field of study (academic experts); And the second group, including IT specialists working in the country's banking industry (industry experts), was categorized. The research model is also shown in Figure 2.

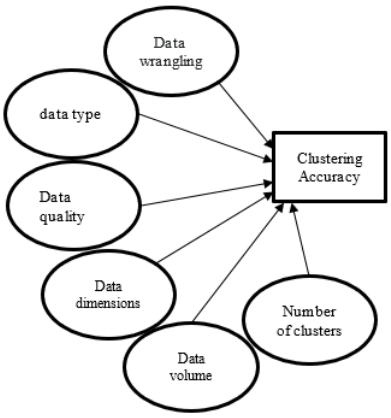

Figure 2: Conceptual model of research

2.2 Implementation of decision tree models

In order to increase the clustering accuracy in big data using data mining techniques such as C&R Tree, Chaid, and C5.0 to evaluate data wrangling indicators, data type, data quality, data dimensions, data volume and number of

clusters as Inputs and clustering accuracy index are treated as outputs.

2.2.1 C&R tree method

In the first step of this method, the data is first called in Excel in SPSS Modeler software. The data are then randomly divided into two categories of training and experiment with 70 and 30 percent, respectively. Then, by specifying data wrangling indicators, data type, data quality, data dimensions, data volume and number of clusters as inputs and clustering accuracy index as output are examined by C&R Tree method. The tree was formed with a maximum depth of 3, the results of the importance of each indicator are shown in Figure 3:

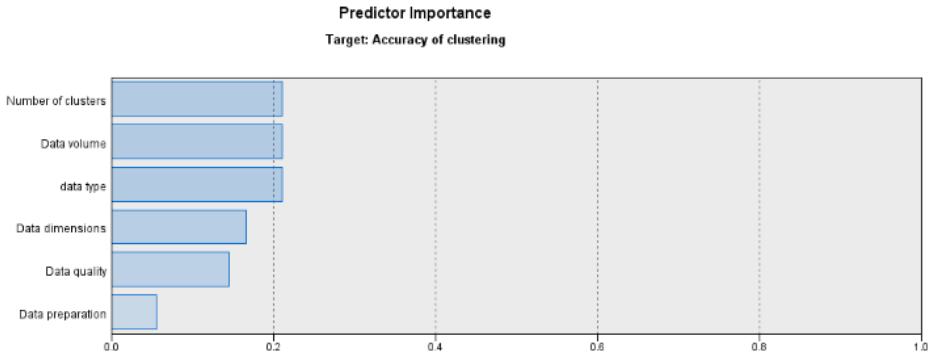

Figure 3: Demonstrate the importance of each variable on *e*-marketing based on the C&R Tree method

2.2.2 CHAID tree method

In the first step of this method, the data is first called in Excel in SPSSModeler software. The data are then randomly divided into two categories of training and experiment with 70 and 30 percent, respectively. Then, by specifying data wrangling indicators, data type, data quality, data dimensions, data volume and number of clusters as inputs and clustering accuracy index as output are examined by C&R Tree method. The tree was formed with a maximum depth of 3, the results of the importance of each indicator are shown in Figure 4:

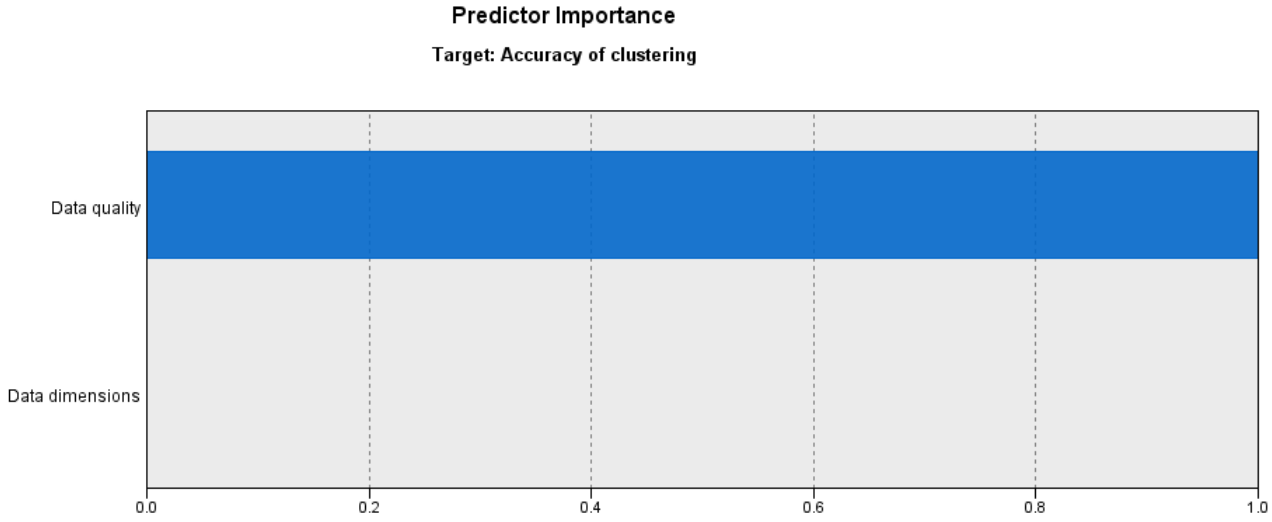

Figure 4: Demonstrate the importance of each of the variables on *e*-marketing based on the Chaid Tree method

2.2.3 C5.0 tree method

In the first step of this method, the data is first called in Excel in SPSSModeler software. The data are then randomly divided into two categories of training and experiment with 70 and 30 percent, respectively. Then, by

specifying data wrangling indicators, data type, data quality, data dimensions, data volume and number of clusters as inputs and clustering accuracy index as output are examined by C5.0 Tree method. The tree was formed with a maximum depth of 3, the results of the importance of each index are shown in Figure 5:

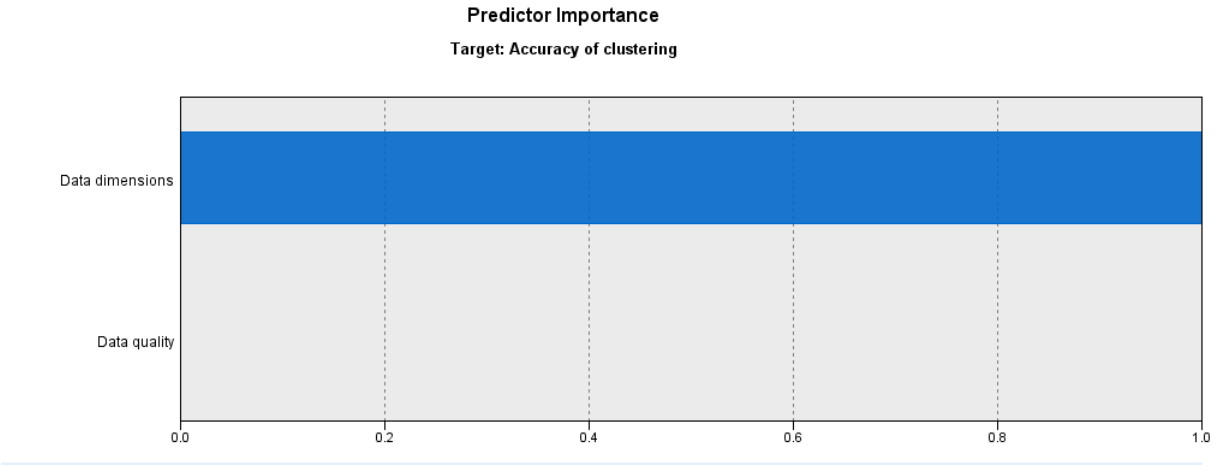

Figure 5: Demonstrate the importance of each variable on *e*-marketing based on the C5.0 Tree method

3 Results

Database, in its general sense, is a collection of information with a regular and organized structure. In this sense, the simple storage of information in a file can also be considered as a kind of database. But in a specific sense, a database is a collection of this information stored in a format that can be read and accessed by electronic devices. The following are some of the academic definitions of this concept. A database is a collection of logically related data (and descriptions of this data) designed to meet an organization's information needs.

The database used in this study was collected in consultation with university professors in the field of study (academic experts) and including IT professionals working in the country's banking industry (industry experts). So that the number 1 indicates very low, the number 2 indicates low, the number 3 indicates normal, the number 4 indicates high and the number 5 indicates very high for each indicator.

Important indicators to increase the accuracy of clustering in large data volumes are:

- Data wrangling
- Data type
- Data quality
- Data dimensions
- Data volume
- Number of clusters

Based on the results of decision tree models, data wrangling indicators, data type, data quality, data dimensions, data volume and number of clusters as inputs at very good, good, normal, bad and very bad levels and the output index of the clustering accuracy model at the levels Very good, good, normal, bad and very bad are presented using a fuzzy inference system. Each of the inputs at different domains in different conditions is specified by a triangular membership function at separate levels. Also, the output criteria are specified in different domains in different conditions with a triangular membership function at separate levels. The membership functions of the input and output variables are shown in Figures 6 to 12.

There are many fuzzy inference models. The most famous of them in engineering sciences is Mamdani fuzzy model. In this research, Mamdani implication algorithm has been used. This model is preferred to other existing models due

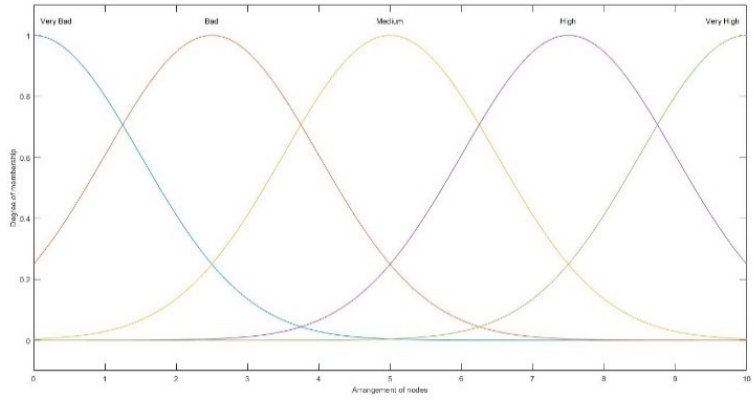

Figure 6: Display membership of, data wrangling input variables

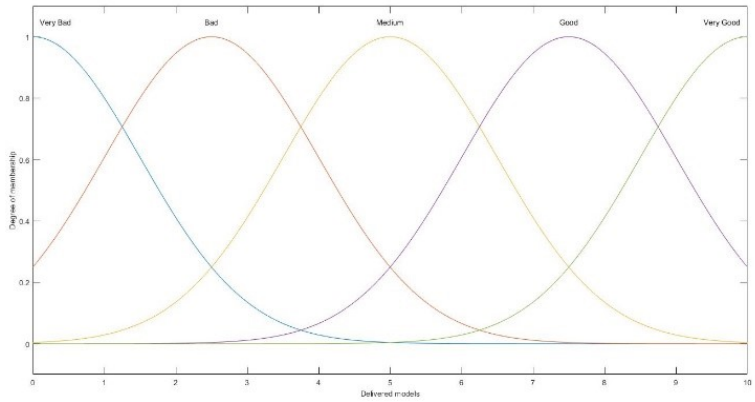

Figure 7: Display membership functions of data type input variables

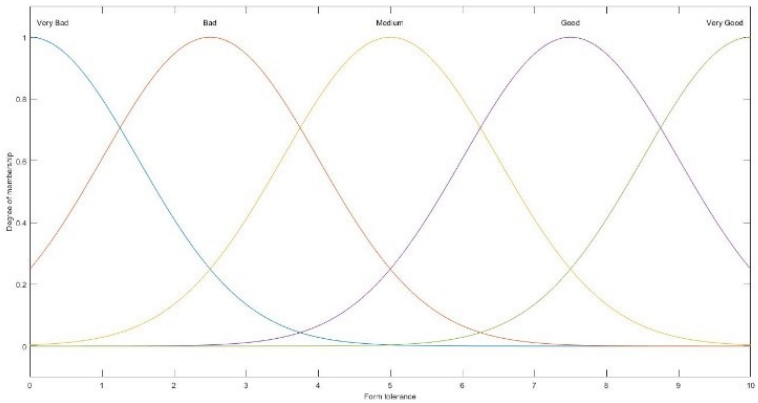

Figure 8: Display membership functions of data quality input variables

to its general acceptance and ease of use. Fuzzy set output membership functions in Mamdani fuzzy inference must be non-fuzzy. This method increases the efficiency of the non-fuzzy process by reducing the required computations. Types of Disposal Methods Included Types of Disposal Methods include Centroid of area, Bisecter of area, Smallest of maximum, Largest of maximum, Mean of maximum, Weighted average (WA), Total Weight, which generally has the most applications of COA and WA. In Mamdani fuzzy model used in this research, a combination of area center of

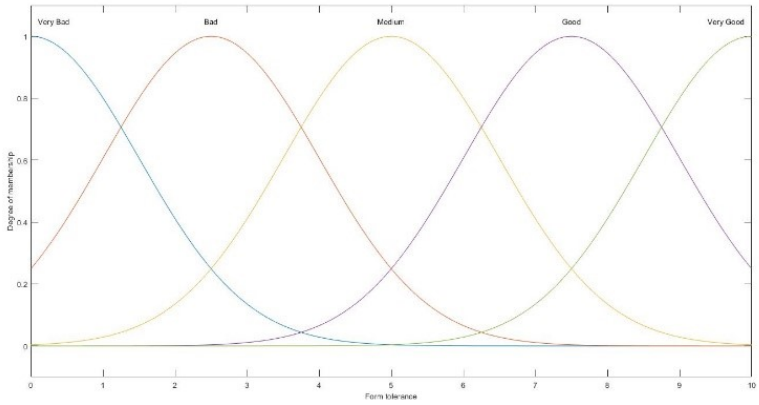

Figure 9: Display membership functions of data dimension input variables

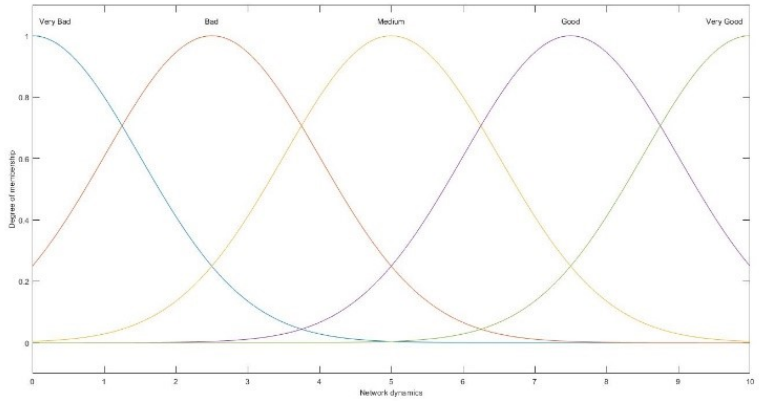

Figure 10: Display membership functions of data volume input variables

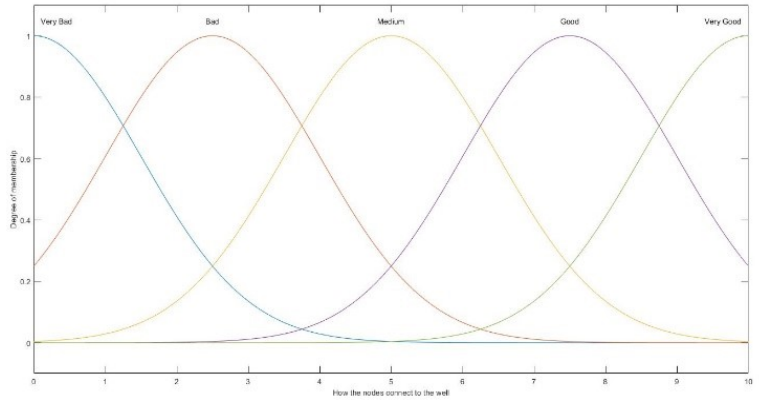

Figure 11: Display membership functions of variable input cluster number

gravity operators has been used. The weight of each of the criteria is 1 and is considered from the conditional phrase "and". Fuzzy inference systems include 30 rules for forming a fuzzy inference system. By applying the membership functions and rules of each fuzzy inference system, the results of the systems can be seen in Figure 13.

The results of the fuzzy inference system can be shown in graphical representation and enter input and output analysis quantitatively. Inputs and outputs were examined by logging in and displaying its output, the results of which

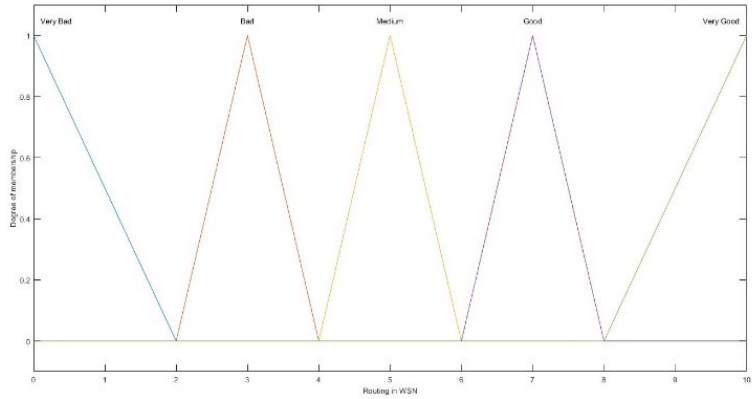

Figure 12: Display membership variables of variable output clustering accuracy

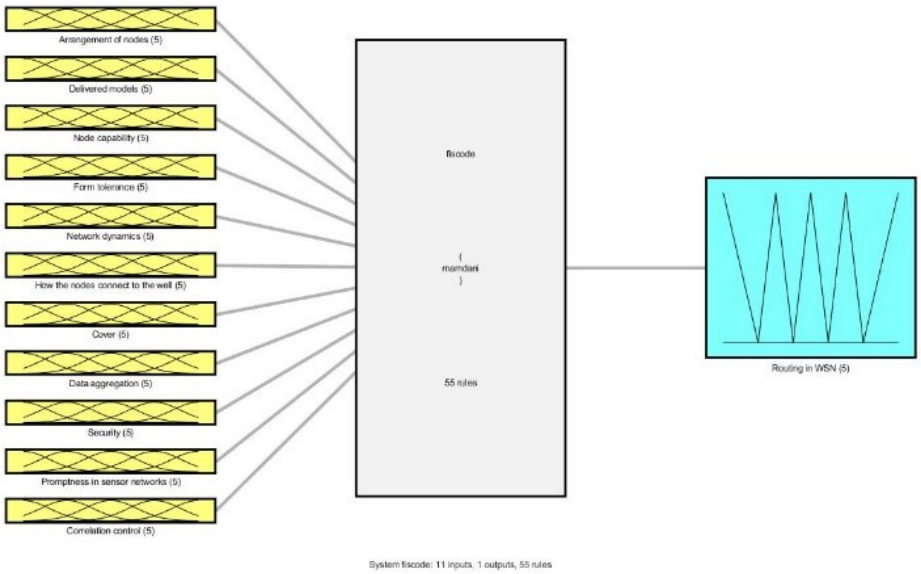

Figure 13: Fuzzy inference system provided for accurate clustering

in Table 1 show that the system has been implemented. Also, the 3D representation of input variables and output variables is performed in Figures 14 to 16.

Table 1: Outputs in a fuzzy inference system

Qualitative clustering accuracy	Slightly clustered accuracy	Data wrangling	Data type	Data quality	Data dimensions	Data volume	Number of clusters
Bad	2.9156	3	2	3	1	1	2
Normal	4.1685	4	9	5	2	3	3
Normal	3.0522	3	5	4	2	2	6
Normal	4.6895	1	2	8	3	4	8
Good	6.7965	9	5	9	6	8	7

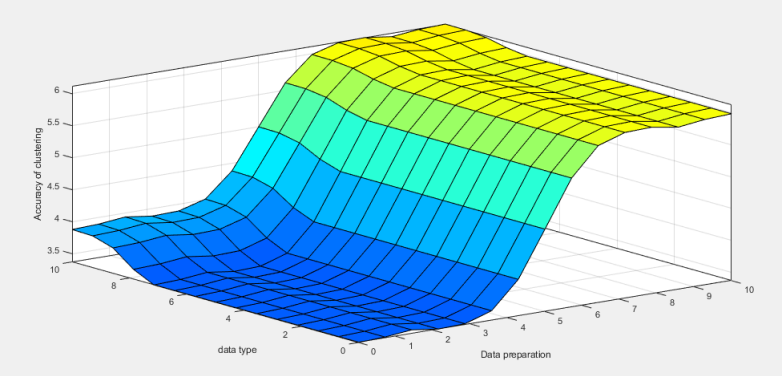

Figure 14: Diagram of data preparation changes and data type and clustering accuracy

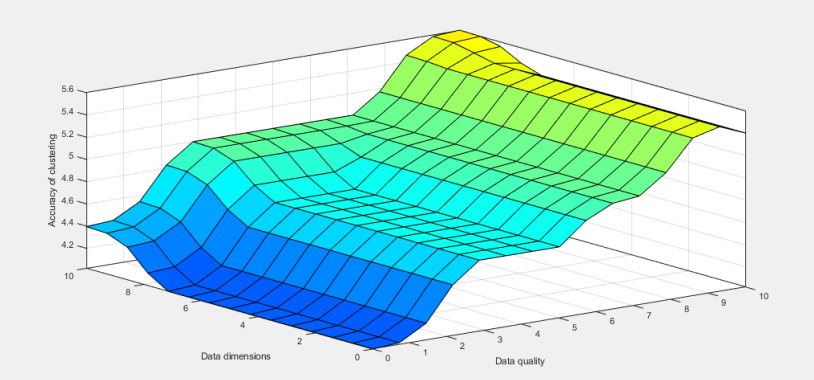

Figure 15: Diagram of data quality changes and data dimensions and clustering accuracy

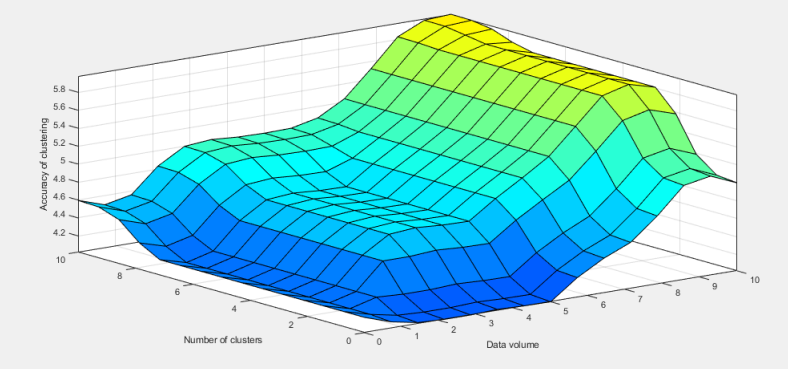

Figure 16: Diagram of data volume changes and number of clusters and clustering accuracy

4 Conclusions

By reviewing the conducted researches, the factors related to increasing the clustering accuracy in big data were determined. By collecting data on 24 indicators to help expert's database record was collected. Using data mining techniques such as C&R Tree, Chaid and C5.0, data preparation indicators, data type, data quality, data dimensions, data volume and number of clusters which results in the following method:

- C&R Tree method: number of clusters
- Tree CHIAD method: data quality
- Tree Method C5.0: Data Dimensions

Findings indicate that there is evidence of good performance of the proposed models. According to Table 2, the

accuracy of the results obtained from C&R Tree, Chaid and C5.0 methods is in both training and testing modes. The ranking of the methods in terms of accuracy can be arranged as follows:

1. C&R Tree
2. C5.0 Tree
3. CHAID Tree

Table 2: Investigating the accuracy of models

		Training stage	**Test stage**
C&R Tree method	Correct	%84.21	%100
	wrong	%15.79	0%
Tree CHIAD method	correct	%89.47	80%
	wrong	%10.53	20%
Tree C5.0 method	correct	%94.74	100%
	wrong	%5.26	0%

According to the modeling of the fuzzy inference system based on the results of the decision tree model related to 6 inputs and outputs, taking into account the relevant rules, the result of the status of each of the inputs and outputs can be examined in Figures 17 to 22.

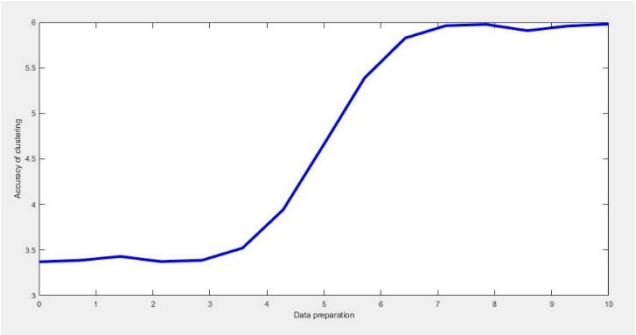

Figure 17: Check data wrangling and clustering accuracy

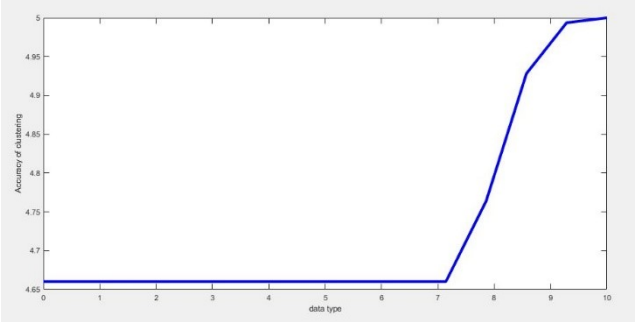

Figure 18: Check data type and clustering accuracy

This study uses data mining techniques such as C&R Tree, Chaid and C5.0 to study the development of FCM method to increase clustering accuracy in high volume data and related factors such as data preparation indicators, data type, Data quality, data dimensions, data volume and number of clusters were evaluated as inputs and clustering accuracy index was evaluated as output. Then, with the help of these results, the rules for forming a fuzzy inference system were determined and by explaining the membership functions of the decision model, it showed what effect each input index has on the output index.

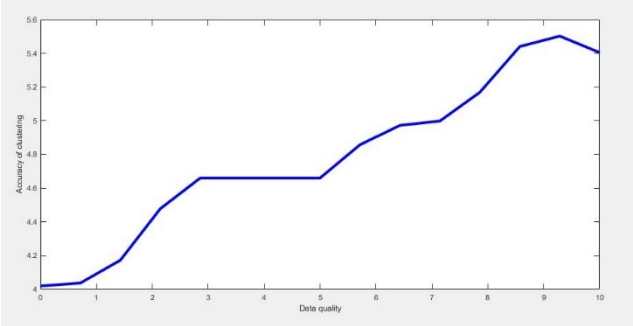

Figure 19: Check data quality and clustering accuracy

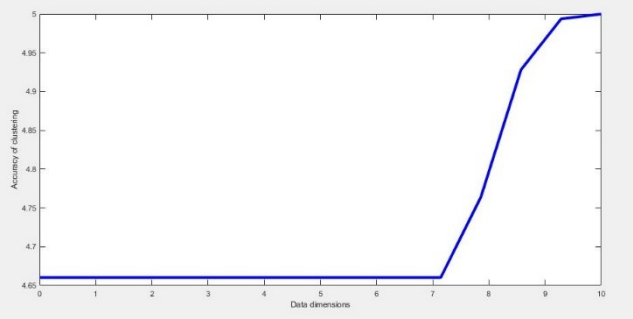

Figure 20: Check data dimensions and clustering accuracy

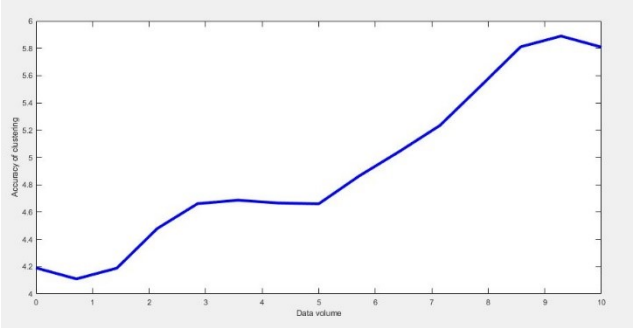

Figure 21: Check data volume and clustering accuracy

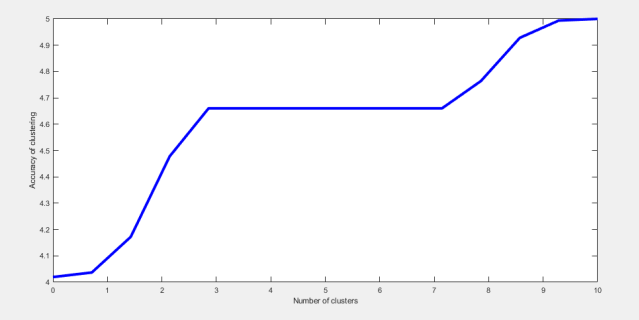

Figure 22: Investigate the relationship between the number of clusters and the accuracy of clustering

References

[1] T. Ahmed, Z. Xiaofei, Z. Wang and P. Gong, *Rectangular array of electromagnetic vector sensors: tensor modelling/decomposition and DOA-polarisation estimation*, IET Signal Process. **13** (2019), no. 7, 689–699.

[2] D. Angrave, A. Charlwood, I. Kirkpatrick, M. Lawrence and M.H.R. Stuart, *analytics: why HR is set to fail the big data challenge*, Human Resource Manag. J. **26** (2016), no. 1, 1–11.

[3] M.R. Bendre and V.R. Thool, *Analytics, challenges and applications in big data environment: a survey*, J. Manag. Anal. **3** (2016), no. 3, 206–239.

[4] S. E. Bibri, *The IoT for smart sustainable cities of the future: An analytical framework for sensor-based big data applications for environmental sustainability*, Sustain. Cit. Soc. **38** (2018), 230–253.

[5] F. Bu, *An efficient fuzzy c-means approach based on canonical polyadic decomposition for clustering big data in IoT*, Future Gen. Comput. Syst. **88** (2018), 675–682.

[6] J. Dekhtiar, A. Durupt, M. Bricogne, B. Eynard, H Rowson and D. Kiritsis, *Deep learning for big data applications in CAD and PLM–Research review, opportunities and case study*, Comput. Ind. **100** (2018), 227–243.

[7] M. Hajeer and D. Dasgupta, *Handling big data using a data-aware HDFS and evolutionary clustering technique*, IEEE Trans. Big Data **5** (2017), no. 2, 134-147.

[8] B. Jan, H. Farman, M. Khan, M. Imran, I. U. Islam, A. Ahmad, ... and G. Jeon, *Deep learning in big data Analytics: A comparative study*, Comput. Electric. Engin. **75** (2019), 275–287.

[9] J. Li, Z. Lu, W. Zhang, J. Wu, H. Qiang, B. Li and P.C. Hung, *SERAC3: Smart and economical resource allocation for big data clusters in community clouds*, Future Gen. Comput. Syst. **85** (2018), 210–221.

[10] D. Liu, L. Ma, X. Liu, H. Yu, H. Tan, X. Zhao, Y. Zhao and G. Lv, *Research on key issues of data integration technology in electric power system in big data environment*, IEEE 9th Int. Conf. Commun. Software Networks, 2017, pp. 1368–1372.

[11] Z. Qingchen, T.Y. Laurence, C. Zhikui and L. Peng, *A survey on deep learning for big data*, Inf. Fusion **42** (2018), 146–157.

[12] M.J. Rezaee, M. Jozmaleki and M. Valipour, *Integrating dynamic fuzzy C-means, data envelopment analysis and artificial neural network to online prediction performance of companies in stock exchange*, Phys. A: Statist. Mech. Appl. **489** (2018), 78–93.

[13] N. Sajadfar and Y. Ma, *A hybrid cost estimation framework based on feature-oriented data mining approach*, Adv. Engin. Inf. **29** (2015), no. 3, 633-647.

[14] G. Suciu, V. Suciu, A. Martian, R. Craciunescu, A. Vulpe, I. Marcu, Simona Halunga and O. Fratu, *Big data, internet of things and cloud convergence–an architecture for secure e-health applications*, J. Med. Syst. **39** (2015), no. 11, 1–8.

[15] S.F. Wamba, S. Akter, A. Edwards, G. Chopin and D. Gnanzou, *How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study*, Int. J. Prod. Econ. **165** (2015), 234–246.

[16] Y. Yang, E.W. See-To and S. Papagiannidis, *You have not been archiving emails for no reason! Using big data analytics to cluster B2B interest in products and services and link clusters to financial performance*, Ind. Market. Manag. **86** (2020), 16–29.

[17] Q. Zhang, L.T. Yang, A. Castiglione, Z. Chen and P. Li, *Secure weighted possibilistic c-means algorithm on cloud for clustering big data*, Inf. Sci. **479** (2019) , 515–525.