AN INNOVATIVE ARTIFICIAL INTELLIGENCE APPROACH USING DATA MINING CLUSTERING ALGORITHM

[1]**Dr. A Ramamurthy**, Professor, Dept. of CSE, DNR College of Engineering and Technology, Bhimavaram, A.P, India
[2]**Dr. BVS Varma**, Professor, Dept. of CSE, DNR College of Engineering and Technology, Bhimavaram, A.P, India

ABSTRACT: Recently huge amount of data is present in internet as the network technology and the information technology has been developing rapidly. But there is lack of knowledge which becomes a serious problem. The Data mining in the cloud combines application of conventional data mining under the cloud computing. One of the most popular and widely used algorithms in this cloud data mining is the clustering algorithm. The clustering algorithm maps each and every data to one group which is called clusters and hence forms a clean partition of the specified data. In this paper an innovative artificial intelligence approach using Data mining clustering algorithm is introduced. One of the most popular unsupervised clustering algorithms is FCM algorithm. The FCM algorithm is developed based on the fuzzy entropy function. In this, Probability Based Matching (PBM) index as well as F-measure method is used to validate the clustering results. Because there is requirement in FCM algorithm to define the number of clusters and to define the different cluster values corresponding to different fuzzy partitions. From the results it can be shown that the introduced fuzzy c-mean algorithm with fuzzy entropy can achieve better performance compared with the traditional FCM algorithm and the optimum number of clusters can be determined automatically.

KEYWORDS: Data mining, Cloud Computing, Fuzzy K-means Clustering algorithm, Artificial Intelligence.

I. INTRODUCTION

At present an incredible speed of data is generated in the social media websites by this modern society because of its increased popularity and fast development. In addition with the ubiquitous social and community activities various types of data is generated constantly with the logical testing, logistic transmission, website access, mobile communication, etc., which indicates that users have entered into a new era of huge increasing of big data. Although this big data in a real time environment isn't just "big" it has an unpredictable data and difficult problems can be solved using the various data structures for technology requirements in data analysis. Big data is simply seemed in a literal point of view as increased data size. An analytical technology is require for this which filter outs the low density or low valued data and then extracts best of data into the high density or high valued data from knowledge [1].

Various types of new technologies, approaches and applications have been generated in the latest decades with the rapid development of informative industry. Those generated new technologies which contribute to the development of big data are of analytics, mass storage and Internet [2]. The data mining has been developed into an inter disciplinary subject over changing and developing from the last few decades in which various disciplines like databases, statistics, pattern recognition, machine learning, parallel computing and artificial intelligence of relevant information are integrated [3]. From the original normal data to today's messy and large amount of data, the objects of data that are examined have been evolved because of the development in the data mining. Thus, research scope has become broadened and the technical requirements became increase.

First a dataset is taken in this and select a set of documents (X). Each of those selected documents (X_i) have 'm' number of features or elements with a m-dimensional vector. Every

one of these features are generally normalized to a uniform scaling before clustering as those m number of features regarding to the each document may have different units. In an m-dimensional elements space every X_i document is considered as a point and a set of points with 'n' number of elements is considered as a set of documents(X) from the geometric point of view. The Number of clusters that are specified prior to the clustering process is used by the FCM algorithm in the analysis of fuzzy clustering method to calculate the coordinates of the cluster centers and the partition matrix. Then for several number of cluster values FCM algorithm is utilized and PBM index as well as F-measure methods are used to calculate the results of clustering validity function for both internal and external performance measurements respectively.

II. DATA MINING AND CLOUD COMPUTING

2.1 Data Mining

The process of extracting inherent, possibly useful and previously unknown information from data is defined as Data mining. The information which is useful to the humans from a huge amount of data is discovered and presented in an understandable format by using the visual, statistical and machine learning methods [4]. The semi automated and automated ways are used in the data mining to determine the significant rules and patterns by examining and analyzing the huge quantity of data. Extracting the huge quantity of data is impossible without automation way [5].

The problem of uncovering the hidden data as well as useful information of the data in large databases with the solved with help of data mining. The Data mining is also called as the KDD (Knowledge Discovery Databases). Prediction and description are the two main goals of this data mining. The future and unknown values of some variables or fields of interest are predicted in the prediction by using the other variables of interests in the database. Whereas the finding the human interpretable patterns which describe the data is focused on the description [6]. Classification, regression, clustering, transform and deviation detections are some of the methods that achieve the prediction and description objectives. A model is included in the dependency modeling that describes the existence of dependency models on two levels such as structural level and quantitative level. The variables which were depending locally on each other is generally specified in a graphical form in a structural level while the strengths of the dependencies are specified by the some numerical scaling in the quantitative level of the modeling [7].

2.2 Cloud Computing

Cloud computing is an umbrella term for anything that includes the provision of Internet hosted services. These services are roughly divided into three types as Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). The cloud computing name was inspired by the symbol of cloud that's frequently used in flowcharts and diagrams to represent the internet [8]. Cloud computing is becoming one of the next industry buzzwords. It brings together the following terms: Computing Grid, Utility Computing, Virtualization, Clustering, etc. The Cloud Computing overlaps to some distribution concepts, grid and utility computing, but its meaning when used properly in the context. The conceptual overlap is due to changes in technology, usage and implementation over the years. The cloud is a self-managed and self-managed resource virtualization. Of course, there is staff to keep the hardware, operating systems and networks in order. From a user or app developer's perspective,

however, this only refers to the cloud. The required services resources for performing the functions with dynamically changing requirements can effectively accessed by the cloud computing [9]. Instead of access requirement from the named resource or specific endpoint, an access from the cloud is required for an application or service developer.

Methods

1. Regression method: A function is learned in this which assigns a data element to a real predictor variable [10]. Predicting the survival possibility of a patient according to series of diagnostic test results, prediction of product consumer demand on a latest product according to the advertisement costs, etc., are some of the examples for this.

2. *Classification Method:* A function is learned by which a data item is classified or assigned to one of predefined classes. Identifying the interested objects automatically over a huge image database and classification of the trends in financial markets are the examples [12].

3. *Clustering:* It is a general descriptive activity that attempts to identify a finite set of categories or groupings that describe data. Exclusive and exhaustive or a wide range of representation can be there for clusters such as hierarchical or overlapping [13]. For example: The discovery of homogeneous sub-populations in marketing database for the consumers.

Transform and Deviation Detection: It focused on identifying the mainly important changes in data by using the values that are measured in prior [14]**.**

Dependency modeling*:* Finding of a model is focused in this in which considerable dependencies can be described between variables. Two levels are there in which these dependences of model exists, they are the structural level and the quantitative level [15].

III. AN INNOVATIVE ARTIFICIAL INTELLIGENCE APPROACH USING DATA MINING CLUSTERING ALGORITHM

Innovative artificial intelligence approach using Data mining clustering algorithm is shown in Figure (1). Documents selection then classifying the selected documents, selecting the appropriate keywords, calculating FCM algorithm and selecting the best scored cluster result with the use of PBM index as well as F-measure are the four major steps presented in the proposed frame work which implement data mining clustering by using Fuzzy C-means algorithm. In this, keywords must be manually defined to represent the documents for the selected documents. Weight for these each defined keyword is calculated by using a weighting scheme called Term Frequency-Inverse Document Frequency (TF-IDF).

Then defined a dataset say X after the completion of TFIDF scheme. Since the original classification has significance in the calculation of the F- measure, it is necessary to define the selected documents in original classification.

The Clustering number "c" is increasing from the c=2 by one when the FCM starts up to the number of M that is defined by the user. Every "c" number is calculated by the both BPM index also the F-measure method. By taking in to the consideration of maximum number of clusters

value “c”, the optimum and best “c” value is selected from the each “c” of those two methods of different cluster validity results.

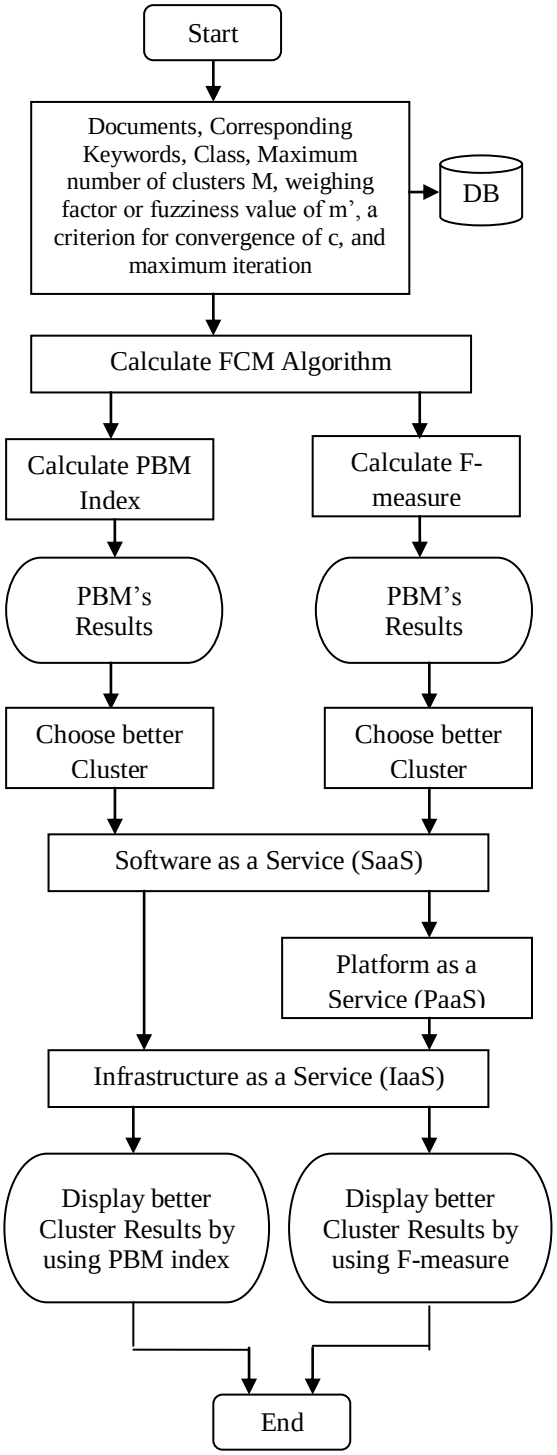

Fig. 1: INNOVATIVE ARTIFICIAL INTELLIGENCE APPROACH USING DATA MINING CLUSTERING ALGORITHM

The Fuzzy c-means objective function is used in this paper with a modified function using a fuzzy entropy H (x) as it is a strictly convex function.

PBM Index

This PBM index measuring method doesn't have external knowledge at all and this is one of the internal performance indicators that depends on the algorithm. The partition with a less number of narrow clusters along with a great partition at least between any two of them is ensured by the maximizing the PBM index.

F-measure

The original classification is compared with the results of clustering by using the F-measure method which was an external performance metric. Let us consider the number of membership values of reference cluster i as n_{ij} in the cluster j, number of membership values as n_j in the cluster j and the number of membership values as n_i in the reference cluster i.

This factor F_c can vary in the range of 0 to 1.

Better clustering is achieved with the larger total F-measure because assigning the clustering to the original classes is more precise. Precision and Recall are computed by counting the number of data elements using the criterion of maximum membership value in each cluster of the fuzzy clusters.

In the cloud architecture, Software as a Service (SaaS) can be considered as a set up for the Data mining. The cloud distributors provide this SaaS facility. This SaaS is utilized in this paper for the introduce clustering frame work in order to implement the data mining clustering according to the data analyst or researcher requirements and parameters. Then the services like Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) of SaaS are used to achieve the data mining security services in the cloud. These three of services together can ensure the frame work to perform the task distribution and scheduling in Data mining properly.

IV. RESULTS

The experiment results for the introduced FCM added with the fuzzy entropy frame work on clustering the document data is presented. The number of documents is selected from the dataset with the 19 number of keywords. Table (1) depicts that six documents selected form the dataset as original classification. Here, various results are obtained by considering the fuzzy c-means partitions in the M=3 number of clusters with the maximum number of 100 iterations and shows various membership values m' as follows.

Table 1: ORIGINAL CLASSIFICATION OF DOCUMENTS

Documents	Original Classification
D1,D2	Crypto
D3, D4	Neural
D3, D4,D5,D6	Document Clustering

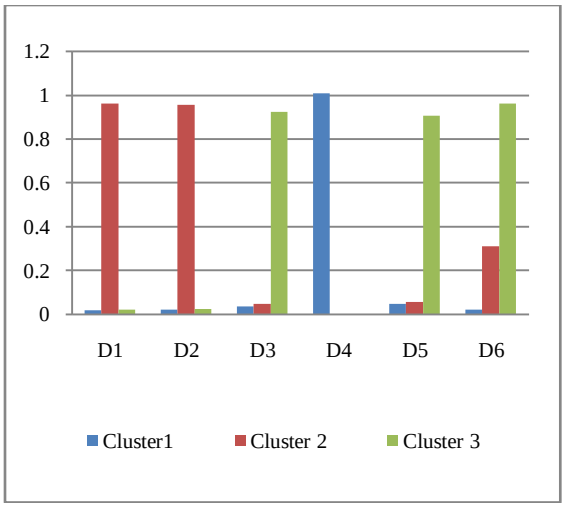

Fig. 2: MEMBERSHIP VALUES OF SIX DOCUMENTS FOR IN THREE CLUSTERS

As depicted in the table (1), both D3 and D4 documents are present in the two classes such as neural in which both are involved in same class and document clustering in which they involve along with the D5 and D6 documents as a class.

The clustering results that obtained are D3, D4, D5 and D6 are regarding to the cluster 1 where as D1 and D2 are regarding as cluster 2 when considering the 2 number of clusters. In this situation it can say that D4 is also involved in cluster 2 as its membership value in it is 0.394. The result with 3 clusters is shown in figure (2).

Even though D4 with the $D3, D5\ and\ D6$ documents has in the same class within the original classification, it also individually include in just one class with a 1 membership value. In a cluster 1, D3 should also include mutually with D4 in comparing with the original classification,

Table 1: PERFORMANCE COMPARISON

SI.NO	Parameter	Traditional FCM Clustering	Innovative artificial intelligence approach using Data mining clustering algorithm
1	Accuracy (%)	90.05	98.93
2	Precision (%)	88.06	98.14
3	Sensitivity (%)	97.65	98.35
4	Specificity (%)	53.84	98.65

The performance of proposed FCM algorithm added with the fuzzy entropy is compared with the traditional FCM algorithm by considering a series of official simple artificial data that contain valid data and noise data. In this dataset number of 20 samples are selected that were numbered

from 1 to 20 in which both the valid and noise data are presented. Then the proposed algorithm is run on this data set. Evaluation is performed for the proposed algorithm in terms of clustering efficacy function, accuracy, precision, sensitivity and specificity are compared with the traditional algorithm as shown in table (2). It was illustrating that proposed fuzzy c-means algorithm added with fuzzy entropy achieves a better performance measures compare to the traditional fuzzy c-means algorithm since it has some deviations because of considering the noise points for clustering as a valid data.

V. CONCLUSION

Clustering analysis has been most commonly used as a significant part of data mining in different fields. There are variety of algorithms for different clustering techniques, each one have their own characteristics and have been utilized in various application areas. This paper proposed data mining clustering using a fuzzy c-means algorithm based on an AI decision mechanism. This Clustering Fuzzy C-means algorithm is a dominant unsupervised technique which constructed and analyzed the data. As initially number of clusters is unknown, usefulness of estimating the optimum number of clusters is proved with the measurements of cluster validity. From the results it can be shown that the better results of clustering calculated using the method of PBM index is differed from the better results calculated with the F-measure due to their individual nature. Then the best clustering results were obtained for proposed Fuzzy c-means algorithm using fuzzy entropy compared to the tradition Fuzzy c-means algorithm by the evaluation of clustering efficacy function, accuracy, precision, sensitivity and specificity.

VI. REFERENCES

[1] Yuan Huang, Qianyu Zhou, Ruixiao Zhao, Yuxing Xiang and Zhe Cheng, “Data Mining Algorithm for Cloud Network Information Based on Artificial Intelligence Decision Mechanism”, IEEE Access, Volume: 8, 2020.

[2] M. Shengdong, X. Zhengxian and T. Yixiang, "Intelligent traffic control system based on cloud computing and big data mining", *IEEE Trans Ind. Informat.*, vol. 15, no. 12, pp. 6583-6592, Dec. 2019.

[3] Fabrizio Marozzo, Domenico Talia and Paolo Trunfio, “A Workflow Management System for Scalable Data Mining on Cloud”, IEEE Transactions on Services Computing, Volume: 11, Issue: 3, 2018.

[4] Z. Xia, X., Wang, X., Sun, Q., et al., "A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data", *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 2, pp. 340-352, Feb. 2016.

[5] L. Jiang, L. D., Xu, H., Cai, Z., et al., "An IoT-oriented data storage framework in cloud computing platform", *IEEE Trans. Ind. Informat.*, vol. 10, no. 2, pp. 1443-1451, May 2014.

[6] O. Rubel, C. Geddes, M. Chen, E. Cormier-Michel and E. Bethel, "Feature-based analysis of plasma-based particle acceleration data", *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 2, pp. 196-210, Feb. 2014.

[7] Temporal Yun Yang and Ke Chen, “Clustering via Weighted Clustering Ensemble with Different Representations”, IEEE Transactions on Knowledge and Data Engineering, Vol: 23, Issue: 2, 2011

[8], Kun-Ta Chuang, Hung-Leng Chen and Ming-Syan Chen, “On Data Labeling for Clustering Categorical Data”, IEEE Trans., Knowledge and Data Engineering, Vol: 20, Issue: 11, 2008.

[9] P.P. Rodrigues, J. Gama and J.P. Pedroso, "Hierarchical Clustering of Time-Series Data Streams", *IEEE Trans. Knowledge and Data Eng.*, vol. 20, no. 5, pp. 615-627, May 2008.
[10] Hichem Frigui and Cheul Hwang, "Fuzzy Clustering and Aggregation of Relational Data With Instance-Level Constraints", IEEE Transactions on Fuzzy Systems, Volume: 16, Issue: 6, 2008.
[11] M.-Y. Yeh, B.-R. Dai and M.-S. Chen, "Clustering over Multiple Evolving Streams by Events and Correlations", *IEEE Trans. Knowledge and Data Eng.*, vol. 19, no. 10, pp. 1349-1362, Oct. 2007.
[12] B.-R. Dai, J.-W. Huang, M.-Y. Yeh and M.-S. Chen, "Adaptive Clustering for Multiple Evolving Streams", *IEEE Trans. Knowledge and Data Eng.*, vol. 18, no. 9, pp. 1166-1180, Sept. 2006.
[13] R. Xu and D. Wunsch,, "Survey of Clustering Algorithms", *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 645-678, May 2005.
[14] R.T. Ng and J. Han, "CLARANS: A Method for Clustering Objects for Spatial Data Mining", *IEEE Trans. Knowledge and Data Eng.*, 2002.
[15] M.-S. Chen, J. Han and P.S. Yu, "Data Mining: An Overview from a Database Perspective", *IEEE Trans. Knowledge and Data Eng.*, 1996.