Indian Journal of Science and Technology, *Vol 9(8), DOI: 10.17485/ijst/2016/v9i8/87971, February 2016*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Survey of Clustering Algorithms for Categorization of Patient Records in Healthcare

D. Narmadha[1*], Appavu alias Balamurugan[2], G. Naveen Sundar[1] and S. Jeba Priya[1]

[1]School of CST, Karunya University, Coimbatore – 641114, Tamil Nadu, India; narmadha@karunya.edu, ggnaveengg@gmail.com, jebapriya@karunya.edu
[2]Department of IT, KLN College of IT, Madurai – 630612, Tamil Nadu, India; app_s@yahoo.com

Abstract

Background/Objectives: This research work provides a survey on the various clustering algorithms such as k-means, K Harmonic means and Hybrid Fuzzy K Harmonic Means (HFKHM) for grouping similar items in large dataset. To improve the accuracy of clustering the large dataset HFKHM is used. **Methods:** The task of analyzing the issues in healthcare databases is extremely difficult since healthcare databases are multi-dimensional, comprising the attributes such as the categorization of tumor, radius, texture, smoothness and compactness of the tumor. This paper presents a related work on the existing clustering algorithms for categorizing the tumors as benign or malignant. Hence clustering algorithms are used to categorize the large dataset based on the diagnosis of the tumor. **Findings:** The efficiency of the various clustering algorithms is compared based on the accuracy and execution time. K means clustering algorithm produces 88% accuracy, 89% accuracy is obtained with the help of K Harmonic Means clustering approach, 90.5% accuracy is achieved using HFKHM clustering approach. **Application:** This model can be an efficient approach for categorizing similar patient records based on the symptoms, treatments and age.

Key words: Clustering, Map Reduce

1. Introduction

The need for efficient data processing and analyzing huge volumes of data is a major challenge in all the areas of research. This limitation can be resolved using Data Mining. Big data can also handle different types of structured and unstructured data such as click streams, audio and video. [1]Progressively unprecedented amount of data is generated from Business Informatics, Social Networks, Meteorology and Health care. Big data analysis has the potential to improve the decision making capability of the health providers through rendering efficient clinical decision support at reduced costs. Big data plays a crucial role in health care. The issues such as similarity of patient records, guaranteeing privacy, safeguarding security, establishing standards and governance are the major challenges to be addressed in the future research.

In recent times, traditional database management systems could not support text analytics in many research areas such as analyzing log records of network, finding purchase pattern of the customers, finding the similarity of patients in terms symptoms, treatments and personal information. Capturing, storing and retrieving the useful information in a timely manner are essential in finding the similarity of patients. Big data helps to produce the solution efficiently. Many researches have been started in the field of finding the patient similarity.

2. Background

Clustering algorithms are broadly classified into distance based method, hierarchical based clustering, partition and probabilistic based methods for grouping similar records.

In the distance based method, similarity is measured in terms of the distance function which is represented as d (i,j). The distance functions operate on various classifications of data such as interval-scaled, Boolean, categorical,

*Author for correspondence

ratio and ordinal variables. In the hierarchical based clustering, there are two broad classifications of methods such as agglomerative and divisive. In the partition based method[2], data is divided into proper subsets and recursively goes through each subset and relocate the point between clusters. In the probabilistic based clustering, data is picked from mixture of probability distribution. Mean and variance are used as parameters for cluster.

Figure 1 shows the pictorial representation of major classification of clustering algorithms. The clustering algorithms must be able to handle huge volumes of data, data of different varieties such as numerical or categorical.

According to the work proposed in[3], another way to rectify optimization process by soft assignment of points to different clusters with appropriate weights, rather than by moving them decisively from one cluster to another. The weights take into account how well a point fits into recipient clusters. This process is called harmonic means. In[4]**,** found the selection of initial centroids is the major problem in k-means algorithm. They also found the same problem in the distributed clustering algorithm based on k-means clustering. Hence a particular factor in the k-means clustering known as *min ()* is changed into *HA ()*. *HA ()* is the Harmonic average and the dynamic weighs has been altered during each iteration. This made the selection of centroids less sensitive and the execution time is dramatically less compared to k-means algorithm.

The stochastic algorithm discussed in[5] is suitable for large samples of high dimensional and fast processing of large datasets. Early k-means algorithm is not suitable for high dimensional dataset. Hence data is increased rapidly on for coming year data driven approach has been followed. Data partitioning is very important on high dimensional large dataset. Sequential version of k-means algorithm is used for fast processing of large dataset.

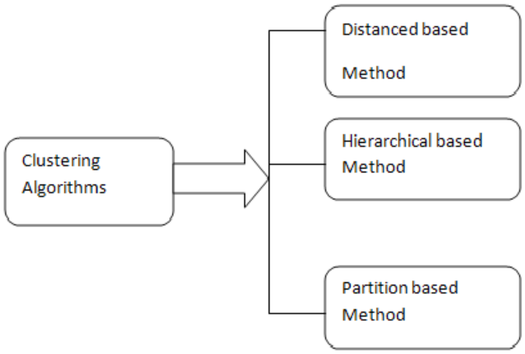

Figure 1. Classification of clustering algorithms.

In[6] a technique of clustering is discussed with two phases. In the first phase, demographic clustering algorithm is done for data set cleaning and creates new patterns using IBM I-Miner. In the second phase data profiling, developing clusters and identify the high value low risk customers is performed.

According to the work proposed in[7], GAKM (Genetic Algorithm based K Means) a hybrid method is introduced that combines a Genetic Algorithm (GA) and k-means algorithm. The major function of GAKM is to determine the optimal weights of attributes and cluster center classification is effective using this algorithm.

As discussed in[8], the problem of selecting initial points in a large dataset can be reduced effectively using canopy clustering algorithm. This proposed algorithm is mainly suitable for high dimensional large dataset. The key idea of applying new mechanism of canopy clustering is to perform clustering in two stages, first a rough and quick stage that divides the data into overlapping subsets known as "canopies" then added rigorous final stage in which expensive distance measurements are only made among points that occur in a common canopy.

In this the first stage is nearly inexpensive methods for finding the center data point. Once the canopies are built using the approximate distance measure, the second stage completes the clustering by running a standard clustering algorithm using a rigorous distance metric. The execution time is relatively less compared to k-means algorithm. In[9]**,** the proposed system comprises of five stages such as Training, Noise Elimination, Learning, Classification and Testing Stages. From the training document, the clustering scheme is built each category and a fuzzy relation is used to measure the similarity between a test document and a category. This relation is called fuzzy term-category relation, where the set of membership degree of words to a particular category represents the cluster prototype of the learned model. Based on this relation, the similarity between a document and a category's cluster center is calculated using fuzzy conjunction and disjunction operators.

According to the method discussed in[10]**,** a credibilistic clustering is used instead of possibilistic clustering. The possibility measures are for possibilistic clustering and credibility measures are for credibilistic clustering. The possibilistic clustering forms its respective membership, it is vulnerable to noise and this problem coincident with every clusters in a large dataset. Credibility measure does not consider model constraint, noise is reduced effectively and center point are effectively identified. The IPCM clustering method discussed in[11] is based on Improved Possibilistic C Means Clustering. The approach is based on two algorithms such as KHM (K Harmonic Means) and IPCM. Hence it is termed to be as Hybrid Fuzzy K Harmonic Means (HFKHM) algorithm. Noise is major problem in KHM. This Noise factor is highly reduced using HFKHM. This clustering algorithm shows the best accuracy compared to other algorithm. In the fuzzy based method**,** the fuzzy set with the function called as membership function is used. The fuzzy set formation is due to the minimum distance from the cluster center. This algorithm is similar to the k-means algorithm but the Euclidean distance in the k-means is replaced by the fuzzy logic.

In[12] provides an insight about fuzzy logic pattern using the k-means algorithm as a based. Since the k-means algorithm is very effective in terms of execution time. The k-means algorithm preferred Euclidean distance for clustering process. Hence the minimum value is quoted into the cluster. This approach compares the Euclidean distance measure in fuzzy logic with different distance measures such as Harmonic distance; Canberra distance etc. and their outputs are verified.

The hierarchical agglomerative clustering algorithm CURE (Clustering Using REpresentatives) discussed in[13] focuses to attain good scalability. This algorithm features of general significance. Outliers is the major importance and with label assignment stage. It also uses two devices to achieve scalability. The first one is data sampling and the second device is data partitioning in p partitions, so that fine granularity clusters are constructed in partitions first. A major drawback is non suitability of unstructured large dataset.

The hierarchical agglomerative algorithm chameleon discussed in[14]**,** utilizes dynamic modeling in cluster aggregation which is very useful in large dataset. Rest of the working is same as CURE.

In the method discussed in[15]**,** conceptual or model-based approach is used which is based on hierarchical clustering. The model associated with a cluster covers both numerical and categorical attributes and constitutes a blend of Gaussian and multinomial models. This algorithm uses maximum likelihood estimation.

In particular in the year 1970, Kernighan and Lin researched for the refinement in hierarchical clustering by k-way graph partitioning which is found give most likelihood of data objects in large datasets.

In[16] gives a comparison between the diverse kinds of algorithms for clustering big data. The work discusses about partition based, hierarchical based, density based and model based approaches for clustering big data. The three dimensional properties of big data such as volume, velocity and veracity are used to measure the strengths and weaknesses of the algorithm. DENCLUE, BIRCH and OptiGrid are most suitable algorithms for dealing with high dimensional data. The projected space clustering model discussed in the research paper[17] focuses to handle non sequence patterns of data. This approach works on complete set of attributes to form pattern wise clusters. It can handle unequal number of attributes and patterns. However, this approach is unable to handle high dimensional data. In[18] weighted k-means algorithm is developed for identifying the diseases like leukemia, inflammatory, bacterial or viral infection. The performance parameters such as accuracy, error rate, execution time are used to measure the effectiveness of the algorithms. The approach is not able to handle large collection of data.

Table 1 shows the comparison between various fuzzy based clustering algorithms. In this table, fuzzy similarity based self constructing algorithm for feature clustering outperforms fuzzy signature based clustering and fuzzy ontology based clustering approach in terms of accuracy and execution time.

3. Proposed Model

CRM (Customer Relationship Management) is cycle of sales, services and support with customer as major entity. Support sector in the CRM cycle plays a vital role of proper organising problem and attaining solution for it. In each sales and services sector, the problems are handled in the support sector (i.e.) similar to the customer care support provided by each mobile network. As the customer increases, data also started to increase proportionally. For effective functioning of support sector, clustering algorithm is needed to cluster data based on the issues in sales and services sector.

Figure 2 explains the architecture of proposed analytic solution. The proposed method provides an end to end solution for conducting large scale analysis of technical support data using open source with Hadoop as a major platform, components of the Hadoop Extend Ecosystem such as HBase, Hive and clustering algorithm algorithms from the extended mahout library.

Mahout is an open source machine learning library built on top of Hadoop to provide distributed analytics capabilities. Mahout incorporates a wide range of data mining techniques including collaborative filtering,

Table 1. Comparison of various clustering algorithms

Survey	Algorithm	Methodology	Results found
19	Fuzzy based Algorithm	Proposed a fuzzy similarity based self-constructing algorithm for feature clustering. Highly reduces the data dimensionality as each cluster, formed automatically, is characterized by a membership function with statistical mean and deviation. It chooses one extracted feature for each cluster.	Execution time varies for different methods on 20 newsgroups data. Generally for 84 extracted features it needs approx. 17 seconds but according to Distributed Clustering it requires approx. 293. Hence the execution time is higher yet it provides good accuracy
20		Proposed the fuzzy signature based solution using frequent max substring mining because of its language independency and favourable speed and store requirements. Deal with cases to handle complex structure data, to handle overlapping information, to include evolving information easily and to handle missing information.	Signature is implemented in the frequent dataset. This reduces the overlapping in the document
21		Produced an extended fuzzy ontology model Proposed a semantic query expansion technology to implement semantic information query based on the property values and the relationships of fuzzy concepts.	Evaluation is based on the Precision value. But higher accuracy is obtained

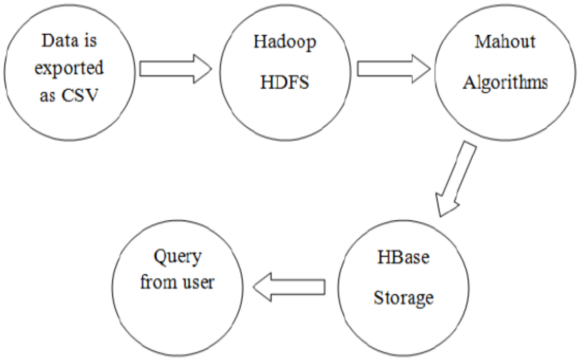

Figure 2. Architecture of proposed analytic solution.

classification and clustering algorithms. Mahout supports a wide variety of clustering algorithms including: k-means, canopy clustering, fuzzy k-means, Dirichlet Clustering and Latent Dirichlet Allocation.

General working of the architecture of proposed analytic solution:

Step 1: Technical support data is exported as Comma Separated Vector (CSV).
Step 2: CSV files are loaded into HDFS daily.
Step 3: Mahout Algorithms are to run and analyse the data.
Step 4: Clustering results are stored in HBase.
Step 5: Users query the clustering results using web interface.

Table 2. Comparison of efficiency of clustering algorithms

Algorithm	Speed of the processor (GHz)	Number of records used	Cluster Accuracy (%)	Execution Time (ms)
	1.36	10076		
KM			88	39901.0
KHM			89	47747.0
HFKHM			90.5	51750.0

4. Experimental Result

The task of analyzing the issues in health care databases is extremely difficult since health care databases are multi-dimensional, comprising the attributes such as the categorization of tumor, radius, texture, smoothness and compactness of the tumor.

The breast cancer dataset is used to compare the similarity algorithm. The similarity algorithm is compared based on accuracy and execution. These performances vary independently based on the system process where the similarity algorithm is performed. Hence the processor used in this dataset is 4 GB RAM with 2.30 GHz as maximum speed. The execution is unstable performance metrics since it varies based on the system processing efficiency. The Table 2 shows the results of three clustering algorithm.

5. Conclusion and Future Work

According to the various surveys, hybrid algorithm and enhanced k-means can obtain less execution time than k-means. K Harmonic means algorithm also provides less execution than k-means but noise is the major drawback in KHM which is overcome by Hybrid Fuzzy K Harmonic Means (HFKHM) algorithm. Table 2 demonstrates the comparison between the implementation of different clustering algorithms such as K-Means, K Harmonic Means, Hybrid Fuzzy K Harmonic Means. The result clearly shows that the HFKHM (Hybrid Fuzzy K Harmonic Means) outperforms the other two algorithms in terms of accuracy.

6. References

1. Fahim AM, Salem AM, Torkey FA, Ramadan MA. An efficient enhanced k-means clustering algorithm. Journal of Zhejiang University Science. 2006 Oct; 7(10):1626–33. ISSN 1009-3095; ISSN 1862-1775.
2. Hung MC, Wu J, Chang JH, Yang DL. An efficient k-means clustering algorithm using simple partitioning. Journal of Information Science and Engineering. 2005; 21:1157–77.
3. Zhang B, Hsu M, Dayal U. K-Harmonic means - A spatial clustering algorithm with boosting. Series Lecture Notes in Computer Science. 2001; 2007:31–45.
4. Thangavel K, Visalakshi NK. Ensemble based Distributed K-Harmonic Means Clustering. International Journal of Recent Trends in Engineering. 2009; 2(1):125–9.
5. Cardot H, Cenac P, Monnez JM. A fast and recursive algorithm for clustering large datasets with k-medians. Computational Statistics and Data Analysis. 2012 Jun; 56(6):1434–49.
6. Rajagopal D. Customer data clustering using data mining technique. IJDMS. 2011 Nov; 3(4):1–11.
7. Jacob SG, Ramani RG. Evolving efficient clustering and classification patterns in lymphography data through data mining techniques. IJSC. 2012 Aug; 3(3):119–32.
8. Kumar A, Ingle YS. Canopy clustering: A review on pre-clustering approach to k-means clustering. IJIACS. 2014 Jul; 3(5):22–9. ISSN 2347 – 8616,
9. Al-Taani AT, Al-Awad NAK. A comparative study of web-pages classification methods using fuzzy operators applied to Arabic web-pages. World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering. 2007; 1(7):2181–3.
10. Kalhori MRN, Zarandi MHF, Turksen IB. A new credibilistic clustering algorithm. Information Sciences. 2014 Sep; 279:105–22.
11. Wu X, Wu BC, Sun JA, Qiu SD, Li X. A hybrid fuzzy K-harmonic means clustering algorithm. Applied Mathematical Modeling. 2015 Jun; 39(12):3398–3409.
12. Das S. Pattern recognition using fuzzy c means. International Journal of Energy, Information and Communications. 2013 Feb; 4(1):1–14.
13. Guha S, Rastogi R, Shim K. CURE: An efficient clustering algorithm for larger databases. ACM. 1998 Jun; 27(2):73–84.
14. Karypis G. A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM Journal on Scientific Computing. 1998 Aug; 20(1):359–92.
15. Chiu T. A robust and scalable clustering algorithm for mixed type attributes in large database environment. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2001. p. 263–8.
16. Fahad A, Alshatri N, Tari Z, Alamri A, Khalil I, Zomaya AY, Foufou S, Bouras A. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE Transactions on Emerging Topics in Computing. 2014 Sep; 2(3):267–79.
17. Gokila S, Ananda Kumar K, Bharathi A. Modified projected space clustering model on weather data to predict climate of next season. Indian Journal of Science and Technology, 2015 Jul; 8(14):1–5.
18. Vijayarani S, Sudha S. An efficient clustering algorithm for predicting diseases from hemogram blood test n Journal of Science and Technology. 2015 Aug; 8(17):1–8.
19. Jiang JY, Liou RJ, Lee SJ. A fuzzy self-constructing feature clustering algorithm for text classification. IEEE Transactions on Knowledge and Data Engineering. 2011 Mar; 23(3):335–49.
20. Wong KW, Chumwatana T, Tikk D. Exploring the use of fuzzy signature for text mining. IEEE International Conference on Fuzzy Systems (FUZZ); Barcelona. 2010 Jul 18-23. p. 1–5.
21. Yang Q, Chen W, Wen B. Fuzzy ontology generation model using fuzzy clustering for learning evaluation. IEEE International Conference on Granular Computing; Nanchang. 2009 Aug 17-19. p. 682–5.