ISSN 2664-4150 (Print) & ISSN 2664-794X (Online)
South Asian Research Journal of Engineering and Technology
Abbreviated Key Title: South Asian Res J Eng Tech

| Volume-4 | Issue-5 | Sept-Oct- 2022 | DOI: 10.36346/sarjet.2022.v04i05.003

Review Article

A Survey of Clustering Methods for Health Care Using Data Mining

T. Aditya Sai Srinivas[1*], Y. Rama Mohan[1], R. Varaprasad[2], G. Mahalaxmi[2], Y. Sravanthi[3], I. Priyanka[1]

[1]G. Pulla Reddy College of Engineering, Andhra Pradesh, India
[2]G. Pullaiah College of Engineering and Technology, Andhra Pradesh, India
[3]Independent Researcher

***Corresponding Author:** T. Aditya Sai Srinivas
G. Pulla Reddy College of Engineering, Andhra Pradesh, India

Article History
Received: 14.08.2022
Accepted: 09.09.2022
Published: 14.09.2022

Abstract: Due to the increasingly expanding medical profession, big data analytics has begun to play a crucial role in advancing healthcare execution and research. It has enabled the collection, management, analysis, and assimilation of huge volumes of unique, structured, and unstructured information generated by contemporary medical service systems. It has provided devices for gathering, directing, analysing, and storing vast quantities of unique, structured, and unstructured data generated by contemporary medicinal administration systems. It produces information in exponentially varied configurations. The medical services division has been well ahead of the curve in adopting this new technology, and it is producing this data at an exponential rate. Consequently, the medical services information contains a substantial amount of information originating from internal and external sources. Payers (claims and cost data), consumers and marketers (patient conduct and feeling data), providers (medical information, government population and general wellbeing information), developers (Pharmacy and therapeutic device research and development), and researchers and scientists (academic and independent) are among the information sources. Because data isn't always the same, each of these data storage facilities is also becoming more diverse, as shown by the four Vs: volume, velocity, variety, and veracity.

Keywords: Big Data, Medical services, and Health care.

1. INTRODUCTION

According to the World Health Organization, "Big Data" is the growing use of rapidly collected, complicated information needing large storage capacities (terabytes, petabytes, zettabytes, or yottabytes) [1]. Unquestionably, the volume of data produced is continually increasing due to the fact that multi-data is created by a single individual on purpose or by mistake via the consistent use of electronic devices. Specifically, reliance on high-throughput sequencing stages, advancing imaging, and motivation behind care devices, as well as calculating and adaptable prosperity improvements, has paved the way for massive data gathering [2]. The primary challenge posed by big data consists of providing illustrative examples of such vast quantities of multi-organized, cross-platform data. In addition, the social insurance market is constantly evolving in terms of advances in therapeutic and mechanical measurements. This paper provides a concise explanation of the grouping techniques utilised in healthcare as well as the evolution of big data in business.

Applications of Big Data in Healthcare Industry

Despite the late adoption of large data in the pharmaceutical sector, the healthcare industry is privileged to comprehend large data due to the challenges posed by conducting effective research on such large, rapid, and complicated information. In addition, challenges such as data protection, security, quality, and unwillingness to share data slow the adoption of big data analytics [3]. Nonetheless, with a growing population and requests for effective healthcare frameworks for disease prevention, intervention, and control, this is not the case. Current medical practitioners are seeking aid from Big Data. However, it is difficult to transform such large volumes of information into convincing and substantiated data, and there is a pressing need to transform existing devices into effective apparatuses to satisfy the essential requirements of data analysis. The following table describes the use of massive data in healthcare administration [4].

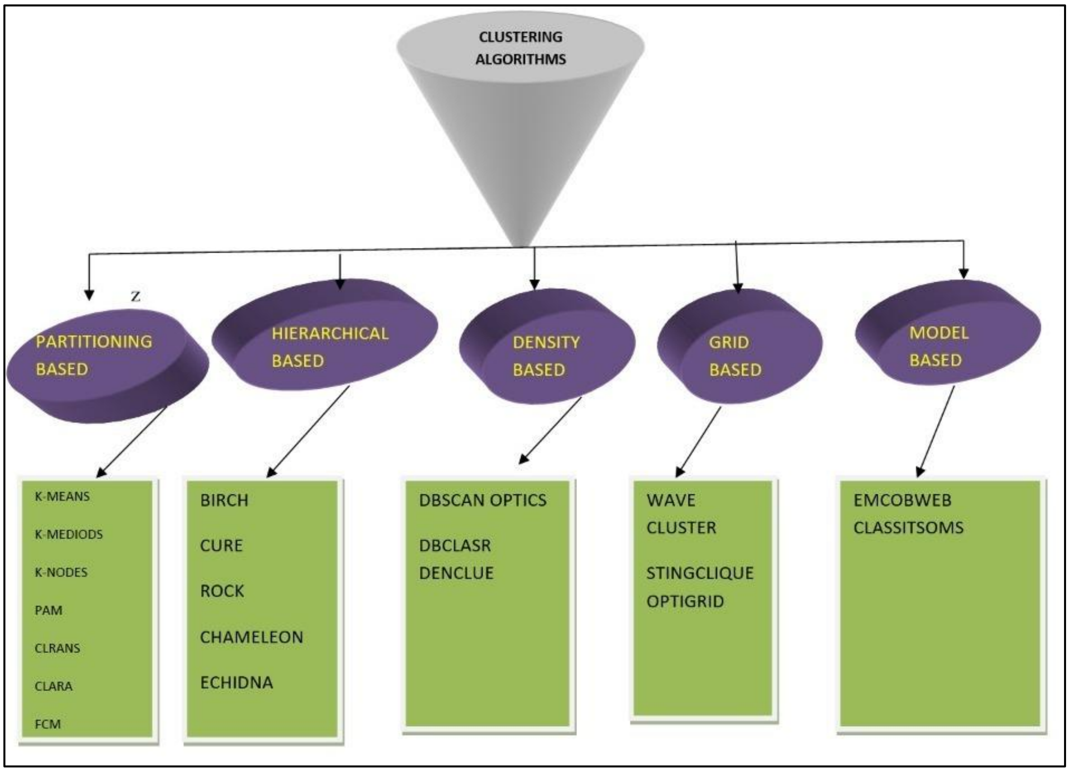
Table 1: Applications of Big data in Healthcare

Source of data	Data type	Example	Data Mining	Advantages	Disadvantages
Electronic Health Records	Structured	Diagnostics, laboratory tests, Medication and auxiliary clinical data encounters and recordings.	Biological information socio-demographic information.	Disease Management Support; health risk assessment, pharmacovigilance, survival rates, therapeutic recommendations, comorbidity prediction of anxiety and suicidal rates real time analysis of mood and behavior when compared to traditional methods.	Electronic health records data misinterpretation, poor signals, errors in recording.
Mobile phone	Text, audio, video		Biological, geographical, socio-demographic information	Prediction of type 2 diabetes; develop interventions for lifestyle modification	limited evidence on how mobile based interventions improve health status and health behaviour
Pharmacogenomics	genomic, proteomic, and metabolic datasets	Whole Genome Sequencing	Genetic information	Predict complex diseases like cancer, TB; customized treatment; drug repositioning; predict drug combination; predict mechanism of action; drug development	Challenges of integration and manipulation of diverse genomic data on big data
Smart phones with sensors	Real time analysis	Heart rate and heart rate variability, energy expenditure, BP, pulse transmission time.		Real Time CARDIO VASCULAR DISEASE Prediction Detection of Parkinson‟s disease human activity recognition	**Advantages:** Better interaction with patient and healthcare providers, personalized treatment, high responsiveness
Wearable, Implantable, and Ambient Sensors	Real time analysis			Prediction of acute and chronic disease; rehabilitation self-monitoring capture critical events and stream data of health information; continuous monitoring of blood glucose levels,	**Advantages**: It continuously monitors and does not affect daily activity, portable, wearable, implantable
Magnetic resonance imaging/positron emission /tomography(MRI)/pet; computed tomography (ct/pet)	Image processing	proteins, cells, tissues, and organs	Biological information	Neuro-imgaing studies	Challenge of collating data from image for prediction, diagnosis, and treatment of diseases. Requires effective and optimized querying systems to reduce
Telemedicine	Unstructured	Discussions, sharing	Biological, geographical,	Identification of epidemic diseases;	Need for data compression
Social media and internet searches	Unstructured	searches	Biological, geographical, sociodemographic information	Public health surveillance; prediction of environment related diseases like asthma; air quality data prediction of psychological state such as hostility and stress	Technical challenges in collating data from social media, privacy and security concerns

Source of data	Data type	Example	Data Mining	Advantages	Disadvantages
				levels; prediction of heart disease based on geographical location.	
Remote sensing technology and geographic information systems	Sensor networks	Weather prediction	Geographical information	Predict impact of pollution on human health; prediction of heat-cold stress related mortality; infectious disease surveillance.	Remote sensing technology and geographic information systems

2. Machine Learning Algorithms in Healthcare

A few bunching calculations are in presence for recognizing comparative data objects. Previously in healthcare [5], generally three techniques that was largely preferred in the field of data mining, logistic regression analysis, Artificial Neural Networks (ANN) and decision tree [6].



Sumana and Santhanam's [7] research developed a high breadth model employing K-means clustering, Best First Search (BFS), and Connection-based Element Selection (CFS). More than 2000 instances of breast cancer, liver disease, heart disease, and diabetes were obtained from the UCI data warehouses [8, 9]. Course k-implies bunching with BDS and CFS demonstrated precision of at least 95% when compared to 12 single models, according to the research. Research by Bansal *et al.,* suggested that k-implies group-based data allocation for cancer forecasting.

The Cancer dataset was obtained and loaded into MATLAB for analysis [10]. After the dataset is plotted, the group k- implied technique is used to group the dataset, and then Support Vector Machine (SVM) [11] is used to arrange the dataset. The standardisation approach was implemented to produce superior quantities. The generated demonstration was compared to the previous k-implies technique, and it was determined that the accuracy rate increased to 92.86 percent, while the execution time increased from 8.5 seconds to 8.5. When bigger datasets with a lot of records are available, the level of exactness may go down [12].

In anticipation of diseases, Deotare *et al.,* conducted research employing pre-processing methods, fuzzy logic, and the K- Means clustering algorithm. The dataset consisted of printed representations of indications obtained from the UCI data vault and mined with KCA to extract top terms. The proposed display's suitability was evaluated on a Java-based Windows PC with an Apache Tomcat server. The investigation concluded that the proposed display had a high rate of accuracy. However, the model was not evaluated for precision, specificity, or influence.

According to research conducted by Biomed Research International, medical image analysis encompasses multiple domains, including image acquisition, arrangement, remaking, enhancement, transmission, and pressure. New technological advancements have increased the resolution, measurement, and accessibility of multimodal images, which has led to an increase in the accuracy of analysis and therapy modification. Nonetheless, combining therapeutic images with different modalities or other medical data is a possible opportunity. To breakdown this data in a clinical setting, new logical structures and methods are required. These strategies address a variety of issues, gaps, and difficulties, for example, highlights from images that can improve the precision [13] of finding and the capacity to utilise various wellsprings of data to build the precision of conclusion and decrease cost, and enhance the precision of handling techniques, for example, therapeutic picture improvement, enrolment, and division to deliver better proposals at the clinical level.

Vijayarani and Sudha [14] attempted to predict disease from blood tests using a unique weight-based k-implies method for data mining. The effectiveness of the proposed show was evaluated using the Fluffy C-means and K-implies clustering algorithms. Blood records from the Kovai Output Center, including 524 instances and 13 characteristics, were analysed. Based on the analysis, it was found that the proposed display for different diseases was between 85% and 98% accurate.

Sundar *et al.*, [15] presented the K-Means clustering method for predicting disease from authentic and synthetic datasets. Choice Tree, Innocent Bayes, and Neural Systems were evaluated against the suggested display. In comparison to other models, k-Means grouping demonstrated the highest accuracy at 66%, according to the study. A few data sets were used for evaluation, which was a limitation.

The research by Mirmozaffar *et al.*, [16] examined all data mining grouping algorithms to identify those with the highest accuracy rate. The dataset was obtained from Iran's National Wellbeing Service and consisted of 209 instances and 8 attributes. All data was pre-processed using supervised and unsupervised algorithms, and WEKA was used for acceptance testing. Out of eight cluster algorithms, Sifted Group, Make Thickness-Based Cluster, and Basic K-Means displayed the highest accuracy and used the least amount of time to construct the model.

3. Advantages and Disadvantages of Clustering Algorithm [17]

CLUSTERING AVENUES	TYPE	ADVANTAGES	DISADVANTAGES
Partition based approach and density based avenue	K-means and DBSCAN (HDKA)	k-means is simple and fast, and DBSCAN performs better in the presence of noise	Speed. Robust against noise.
Partition based and rough set avenue	Hybrid fuzzy C means and rough set clustering (HCFR) known as „Rough Set Theory."	Analyzevagueness, uncertainty and incompleteness in information system	Analyze and discover the reliant relationship among data using attributes and concept based on upper and lower approximation of dataset
Partitioning avenue	K-means	• Simple unsupervised learning algorithm • High speed • Measurable • Efficient in large data collection	• Selection of optimal number of cluster is difficult. • Selection of initial Centroids in random. • Applicable only when mean is defined • Unable to handle noisy data
	K-Medoids/PAM (Partitioning around Medoids)	Effective on small datasets • Sensitive to noisy data and outliners	• Ineffective on large dataset as searches for large for best k-medoids among given dataset. • Costly than Kmeans.
	CLARA (Clustering Large Applications)	• Best clustering • Searches for best K-medoids among selected sample dataset • Effectiveness is based on the sample size	Cannot find the best clustering if sampled medoid is not among the best K medoids
	FCM (Fuzzy C-Means)	Simple, Effective	Long computational time

4. CONCLUSION

This paper has surveyed the utilization of huge data examination for handling and breaking down healthcare data. We quickly clarified about various types of strategies in grouping for healthcare, it is perceived from the audit of the writing that the amount of healthcare data is expanding on consistent schedule. The present data overseeing systems for examination are not as potential as data investigation. Accordingly there is a need to centre towards enhancing the execution of huge data investigation. The strategies for huge data examination are fit for preparing, appropriating, catching and dealing with the analysis in a specific frame which makes it simple to get dependable data. To accomplish a more profound comprehension of results gigantic measure of patient-related healthcare data ought to be assessed effectively, which might be connected at the purpose of healthcare for better services.

REFERENCES

1. Gheisari, M., Wang, G., & Bhuiyan, M. Z. A. (2017, July). A survey on deep learning in big data. In *2017 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC)* (Vol. 2, pp. 173-180). IEEE.
2. Belle, A., Thiagarajan, R., Soroushmehr, S. M., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big data analytics in healthcare. *BioMed research international*, *2015*.
3. Jee, K., & Kim, G. H. (2013). Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. *Healthcare informatics research*, *19*(2), 79-85.
4. Sarabu, A., & Santra, A. K. (2021). Human action recognition in videos using convolution long short-term memory network with spatio-temporal networks. *Emerging Science Journal*, *5*(1), 25-33.
5. Dhillon, A., & Singh, A. (2019). Machine learning in healthcare data analysis: a survey. *Journal of Biology and Today's World*, *8*(6), 1-10.
6. Yegnanarayana, B. (2009). *Artificial neural networks*. PHI Learning Pvt. Ltd..
7. Sumana, B. V., & Santhanam, T. (2014, October). Prediction of diseases by cascading clustering and classification. In *2014 International Conference on Advances in Electronics Computers and Communications* (pp. 1-8). IEEE.
8. Varaprasad, R., Ramasubbareddy, S., & Govinda, K. (2022). Event Recommendation System Using Machine Learning Techniques. In *Innovations in Computer Science and Engineering* (pp. 627-634). Springer, Singapore.
9. Reddy, S. R. S., Nalluri, S., Kunisetti, S., Ashok, S., & Venkatesh, B. (2019). Content-based movie recommendation system using genre correlation. In *Smart Intelligent Computing and Applications* (pp. 391-397). Springer, Singapore.
10. Ratan, R., Sharma, S., & Sharma, S. K. (2009). Brain tumor detection based on multi-parameter MRI image analysis. *ICGST-GVIP Journal*, *9*(3), 9-17.
11. Suthaharan, S. (2016). Support vector machine. In *Machine learning models and algorithms for big data classification* (pp. 207-235). Springer, Boston, MA.
12. Reddy, Y. H., Ali, A., Kumar, P. V., Srinivas, M. H., Netra, K., Achari, V. J., & Varaprasad, R. (2022). A Comprehensive Survey of Internet of Things Applications, Threats, and Security Issues. *South Asian Res J Eng Tech*, *4*(4), 63-77.
13. Sarabu, A., & Santra, A. K. (2020). Distinct two-stream convolutional networks for human action recognition in videos using segment-based temporal modeling. *Data*, *5*(4), 104.
14. Vijayarani, S., & Sudha, S. (2015). An efficient clustering algorithm for predicting diseases from hemogram blood test samples. *Indian Journal of Science and Technology*, *8*(17), 1.
15. Devi, T., & Saravanan, N. (2012). Development of a data clustering algorithm for predicting heart. *International journal of computer applications*, *48*(7).
16. Mirmozaffari, M., Boskabadi, A., Azeem, G., Massah, R., Boskabadi, E., Dolatsara, H. A., & Liravian, A. (2020). Machine learning clustering algorithms based on the DEA optimization approach for banking system in developing countries. *European Journal of Engineering and Technology Research*, *5*(6), 651-658.
17. Sisodia, D., Singh, L., Sisodia, S., & Saxena, K. (2012). Clustering techniques: a brief survey of different clustering algorithms. *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, *1*(3), 82-87.