International Journal of Current Trends in Engineering & Technology
www.ijctet.org, ISSN: 2395-3152
Volume: 06, Issue: 01 (January- February, 2020)

Review on Healthcare Data Analysis Based on Unsupervised Machine Learning Algorithms

Ankita Chourasia[1], Harshita Sharma[2]
1IPS Academy, Indore, M.P., India
2 IET, DAVV, Indore, M.P., India
[1]ankita.chourasia29@gmail.com, [2]hsharma@ietdavv.edu.in

Abstract— *the main goal of the data science or data mining process is to extract useful knowledge or information from the big dataset. Data mining is a technique for examining large preexisting databases to generate new information which helps us to determine future trends. It also helps to find a unique pattern and important knowledge from the existing database. Big Data is a new term used to identify the datasets that due to their large size and complexity. Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. The Big Data challenge is becoming one of the most exciting opportunities for the next years. This study paper includes information about what is big data, Data mining, Data mining with big data, Challenging issues and related work. Data mining is a process of deriving required data from a collection of a large dataset and making an analysis of collected data. Clustering is a technique for grouping of a similar dataset in which data within-cluster having similar properties. K-Means is widely using the clustering algorithm in which uniform effect that is producing clusters with relatively uniform size even if the input data have different cluster sizes is the main advantage. In separated clustering different portions are created based on some criteria that are compared with well known K-Mean algorithms given better accuracy. In this thesis to localization of the protein to data sets are taken from well known UCI machine learning repository. Experiments supported the quality information UCI show that the projected technique will turn out a high purity cluster results and eliminate the sensitivity to the initial centers to some extent. Existing clustering approach is applied to the datasets and tries to find which algorithm provides accurate outcome. It is found that K-Mean clustering provides accurate outcomes as compared to and the newly proposed method. To provide the ability to make sense and maximize utilization of such large data amounts of web data for knowledge discovery and decision-making is crucial to scientific advancement.*

Keywords: Data Science, Data Mining, Dataset, Analysis, Clustering, Big data, Supervised Machine Learning Unsupervised Machine Learning.

I. INTRODUCTION

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. Data science is related to data mining and big data [1]. Data science is a concept to unify statistics, data analysis, machine learning, and their related methods. Data Mining (or information Discovery in Databases) describes the big idea of finding "interesting" patterns in large collections of information. There's a large quantity of data available within the data business. This knowledge is of no use until its regenerate into helpful information. It's necessary to research this large quantity of data and extract helpful information from it. Extraction of knowledge isn't the only method we need to perform; data mining so describes the abstract goals of what must be done and depends upon a large range of various techniques to achieve them, like artificial neural networks, cluster analysis different processes such as data cleaning, data integration, information Transformation, data mining, Pattern analysis, and knowledge Presentation. In this process are overcome problem data error, they might be able to use this data in several applications like error detection in large data set, market dataset analysis research, Production control, Science Exploration, etc. Clustering could also be defined as an information reduction tool i.e. used to produce subgroups that are a lot of and a lot of manageable than an individual data point. Clustering is justified as a method used for grouping a large variety of knowledge into significant teams or clusters supported similarity types of objects data. Clusters area unit the teams that know similar on basis of common options and dissimilar to knowledge in different clusters Cluster analysis teams knowledge objects primarily based solely on data found in the knowledge that describes the objects and their relationships. The objects among a bunch are kind of like one another each different and different from the objects in other teams. Cluster analysis is one in every of the key data processing techniques, wide used for several sensible applications in varied rising areas like Bioinformatics. Clustering is an unsupervised methodology that subdivides associate input file set into a desired variety of subgroups so the objects of a similar subgroup are going to be similar (or related) to at least one another and different from (or unrelated to) the objects in different teams [2,3].

1.1 Types of Clustering

Clustering algorithms have many categories like hierarchical-based algorithms, partition-based algorithms, density-based algorithms, and grid-based algorithms. Partition-based clustering is centroid based which splits data points into k partition and each

partition represents a cluster. K-means is a clustering algorithm that is used widely. This technique will be useful in the extraction of useful information using clusters from a huge Database. The overall purpose of the process of data mining is to extract useful information from a huge set of data and converting it into a form that is understandable for further use. For example, Data clustering is a data exploration technique that allows objects with similar characteristics to be grouped to facilitate their further processing [4].

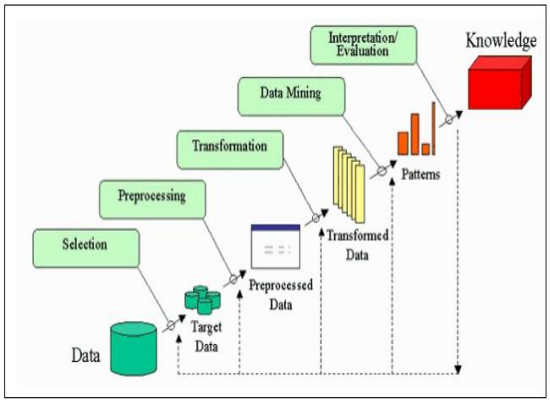

Figure 1 Extract knowledge processing Stages in DS

1.2 Types of Learning Method in Data Mining

In data mining, two learning methods used to mine data i.e. supervised learning and unsupervised learning [5].

Supervised learning: In this learning, data includes together the input and the desired result. It is a fast and perfect learning method. The accurate results are known and are given in inputs to the model during the learning procedure. The neural network, Multilayer perception, Decision tree are supervised models. Unsupervised learning: The desired result is not provided to the unsupervised model during the learning procedure. This method can be used to cluster the input data in classes based on their statistical properties only. These models are for various types of clustering, k-means, distances and normalization, self-organizing maps.

1.3 K-Means Clustering

K-means clustering algorithm is a famous clustering technique. It is used in many areas such as information retrieval, computer vision, and pattern recognition. K-means clustering assigns n data points into k clusters so that similar data points can be grouped. It is an iterative method that assigns each point to the cluster whose centroid is the nearest. Then it again calculates the centroid of these groups by taking its average. Properties of the k-means algorithm [4]: 1. Large data set are efficiently processed. 2. It often terminates at a local optimum. 3. It supports numeric values4. The shape of the clusters is convex [6].

II. LITERATURE SURVEY

Soumi Ghosh et al. [7] proposed a comparative discussion of two clustering algorithms namely centroid based K-Means and representative object-based Fuzzy C-Means clustering algorithms. This discussion is based on the performance evaluation of the efficiency of clustering output by applying these algorithms. The result of this comparative study is that FCM produces a closer result to the K-means but still, computation time is more than k-means due to involvement of the fuzzy measure calculations

Sudipto Guha et al. [8] planned a replacement hierarchical cluster algorithmic rule referred to as CURE that's stronger to outliers and identifies clusters having non-spherical shapes and wide variances in size. This can be achieved in the CURE method by representing every cluster by a particular mounted variety of points that are generated by choosing well-scattered points from the cluster so shrinking them toward the middle of the cluster by a specified fraction. To handle giant databases, CURE employs a mixture of sampling and partitioning. Besides the outline of the CURE algorithmic rule, the author also delineated, the form of options it uses, and why it uses different techniques.

Zomaya et al. [9] present a survey of existing clustering algorithms of different categories (Partitioning-based, Hierarchical-based, Density-based, grid-based and model-based). In their work they established a comparison between five categories with their most representative algorithm; their goal was to find the best performing for Big Data.

Garg et al. [10]. Survey on varied increased K-Means Algorithms". Data mining is defined as a way used to extract and mine the invisible, meaningful info from a mountain of knowledge. Clustering is a very important technique that has been introduced within the space of knowledge mining. Clustering is defined as a technique wont to cluster similar information into a group of clusters supported some common characteristics. K-means is one amongst the popular partition based mostly clustering algorithms within the space of analysis. The impact issue of k-means is its simplicity, high potency, and scalability. However, it additionally includes a range of limitations: random choice of initial centroids, range of cluster K needs to be initialized and influence by outliers. Visible of those deficiencies, this paper presents a survey of enhancements done to ancient k-means to handle such limitations.

Shi Na et al. [11] proposed the analysis of shortcomings of the standard k-means algorithm. K-means algorithm has to calculate the distance between each data object and all cluster centers in each iteration. This repetitive process affects the efficiency of clustering algorithm

Hareesha K. et al. [12] present a changed K-means algorithmic program to enhance the cluster quality and to repair the best variety of clusters. As input variety of clusters (K) given to the K-means algorithmic program

by the user. However, within the sensible state of affairs, it's difficult to repair the number of clusters before. The strategy planned during this paper works for each of the cases i.e. for a famous variety of clusters before additionally as an unknown variety of clusters. The user has the pliability either to repair the variety| of clusters or input the minimum number of clusters needed. The new cluster centers are computed by the algorithmic program by incrementing the cluster counter by one in every iteration until it satisfies the validity of cluster quality. This algorithmic program can overcome this downside by finding the best variety of clusters on the run. The planned approach takes additional machine time than the K-means for larger knowledge sets. It's the foremost disadvantage of this approach.

A.benayed et al. [13] the authors focus on the most popular and most used algorithms in the literature like k-means, they present some comparative work of these algorithms.

Zhang et al. [14] propose a simple and qualitative methodology using k means clustering algorithm to classify NBA guards and used the Euclidean distance as a measure of similarity distance. This work display by using the k-Means clustering algorithm and120 NBA guardsí data. The manual classification of traditional methods is improved using this model. According to the existing statistical data, the NBA players are classified to make the classification and evaluation objectively and scientifically. This work shows that this is a very effective and reasonable methodology. Therefore, based on classification result the guard's type can be defined properly. Meanwhile, the guards function in the team can be evaluated in a fair and objective manner

Sherin et al. [15]. Another recent research presents a general view of data mining algorithms and platforms that can be used in the field of Big Data by discussing different challenges and characteristics.

Wen-Jun Zhang et al. [16]. Adaptive Particle Swarm improvement on Individual-Level". Found out an adaptive particle swarm optimization (PSO) on an individual level. By analyzing the social model of PSO, an exchange criterion supported the variety of fitness between current particles and also the best historical expertise is introduced to keep up the social attribution of swarm adaptively by removing inactive particles. Functions were tested that indicates its improvement within the average performance. An adaptive particle swarm optimization (PSO) on the individual level is conferred. By analyzing the social model of PSO, an exchange criterion supported the variety of fitness between current particles and also the best historical expertise is introduced to keep up the social attribution of swarm adaptively by coming out inactive particles. The testing of 3 benchmark functions indicates it improves the typical performance effectively.

S. Arora et al. [17] discusses some of Big Data mining algorithms to find the most appropriate among them using a comprehensive comparison.

Deepali et al. [18]. Normalization based K means bunch Algorithm. K-means is an economical cluster technique used to separate similar info into groups supported by initial centroids of clusters. Throughout this paper, group action based K-means bunch algorithmic rule (N-K means) is planned. Planned N-K means bunch algorithmic rule applies group action before bunch on the accessible info what is more as a result of the planned approach calculates initial centroids supported weights. Experimental results prove the betterment of planned N-K means bunch algorithmic rule over existing K-means bunch algorithmic rule in terms of complexity and overall performance.

S. Hari Ganesh et al. [19]. "Outlier Detection using increased K-Means clustering algorithm and Weight-based mostly Center Approach". In data processing, their square measure innumerable ways are used to find the outlier by creating the clusters of knowledge then sight the outlier from them. Generally, the bunching technique plays a really necessary role in the data processing. Clustering means that grouping similar information objects along supported the characteristic they possess. Outlier Detection is a very important issue in information mining; significantly it's been wont to determine and eliminate abnormal information objects from given information set wherever outlier is that the information item whose price falls outside the bounds within the sample information could indicate abnormal information. During this work, we've urged a clustering primarily based mostly outlier detection algorithm for effective data processing that uses increased k-means clustering algorithm to cluster the info sets and weight-based center approach. In a planned approach, 2 techniques are combined to expeditiously notice the outlier from the info set. Threshold price is often calculated programmatically by taking the absolute quantity of minimum and most value of a specific cluster. The experimental results demonstrate that increased technique takes the least computational time and concentrates on reducing the outlier that would improve the efficiency of the k-means bunch for achieving the higher quality clusters.

S. Wang et al. [20]. Researchers presents a review of some old algorithms that can handle large data set as Nearest Neighbor Search, Decision Tree, and Neural Network.

Problem Statement: In recent research, a social online medical web forum is a better approach for deciding for disease diagnosis, treatment identification, etc,

based on revives and post comments and helps the society, while that approach under the medical care system. Know recent research, they identify that number of the online site helps the society and provide medical care, so here our aim to develop based modeling for efficient data mining for improving the care of social media users. The existing method k-means clustering method using low accuracy and more flat data retrieve. Verify the accuracy and measure time complexity overcomes this type's problem using our new proposed algorithm (PA).

III. SIMULATION ENVIRONMENT

MATLAB Tool: MATLAB is a software package for high-performance numerical computation and visualization. It provides an interactive environment with hundreds of built-in functions for technical computation, graphics, and animations. The name MATLAB stands for Matrix Laboratory. One of the most features of MATLAB is its platform independence. Once you are in MATLAB, for the most part, it does not matter which computer you are on. In MATLAB the M-files are the standard ASCII text files, with a .m extension to the file name. There are two files of this file: script file and function file. All most programs in write in MATLAB are saved in M-files. Fig-files are binary files with a .fig extension that can be opened again in MATLAB as figures. Such files are created by saving a figure in this format using save or save as option from the File menu or using the save as command in command window-files are compiled M-files with a .p extension that can be executed in MATLAB directly. There are several optional toolboxes are available from developers of MAT-LAB.

IV. PROPOSED APPROACH

In recent research biomedical and health informatics system are based on online web services is very helpful to web users. Because the system provides intelligently knowledge extract from social media and provides improve healthcare to the users in a cost-effective manner. In our approach, we will take the exemplary data of lung cancer from different web sites. The data contain post word, comments, user rank, treatment and side effect of treatment. That input data value converts into the token, after the token creation we inference sentiments of token i.e. positive and negative sentiments, the sentiments passed into the PGA and SOM and k-means clustering method analyze word frequency data that derived from user forum posts, and inference the treatments accuracy and side effect percentage. After the result inference from the SOM modeling, they tune the related post and treatments as well as side effects of treatments for better care of health against lung cancer. Our new proposed algorithm is better as compare to KSOM .because PGA is minimize flat in the dataset and improve the accuracy of retrieval data. Large dataset using healthcare.

IV. CONCLUSION

In this research field of data science, they briefly review and analysis of applications of data mining and this survey are based on various issues of data mining. This paper describes different methodologies and different algorithms used to manage large sets of data. It shows that these algorithms are insufficient to face all the challenges raised by Big Data. Indeed no clustering algorithm can be used to solve all the Big Data issues. Although the parallel classification is potentially very useful for big. Conventional parallel data mining architecture will not provide data mining services in case of network disruption. To simulate an improved innovative and optimal performance analysis based on the medical dataset using data mining techniques to k-means clustering and hierarchical clustering to increase the performance using the medical dataset. Proposed techniques are performance analysis both datasets. Find optimal results based on the medical dataset using SOM and PGA. In experimentation with each the clustering algorithms for used same datasets types with famous clustering algorithms, so that the clustering algorithm with higher accuracy is optimal performance analysis with Find different cluster groups. Here proposed clustering is better as compare to k-mean clustering because more accuracy supported base on minimizing redundancy in the dataset and minimize fault. The results produced were satisfactory in terms of good accuracy. The performance analysis based on the topmost of the clustering algorithm has been compared and analyzed with a number of the existing evolutionary clustering algorithms however this has proved to be more efficient in terms of quality and optimal result. In this experiment, it successfully gets the highest accuracy result to train data using the whole training set. It can be found better outcomes in cross-validation and so many satisfactory results. Data clustering, but the complexity of the implementation of these algorithms remains a great challenge. However, the map-reduce framework can provide a very good basis for the implementation of such parallel algorithms. Generally, to manage a large volume of data while keeping an acceptable resource needs, they have to improve clustering algorithms by reducing, their complexity in terms of time and memory.

REFERENCES

[1]. Cao, Longbing. "Data science: challenges and directions." Communications of the ACM 60, no. 8: 59-68, 2017.
[2]. Berkhin, Pavel. "A survey of clustering data mining techniques." In grouping multidimensional data, pp. 25-71. Springer, Berlin, Heidelberg, 2006.
[3]. Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In KDD, vol. 96, no. 34, pp. 226-231. 1996.
[4]. Verma, Manish, Mauly Srivastava, Neha Chack,

Atul Kumar Diswar, and Nidhi Gupta. "A comparative study of various clustering algorithms in data mining." International Journal of Engineering Research and Applications (IJERA) 2, no. 3 (2012): 1379-1384.

[5]. Kavakiotis, Ioannis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, and Ioanna Chouvarda. "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal 15 (2017): 104-116.

[6]. Likas, Aristidis, Nikos Vlassis, and Jakob J. Verbeek. "The global k-means clustering algorithm." Pattern recognition 36, no. 2: 451-461, 2003.

[7]. Soumi Ghosh, Sanjay Kumar Dubey, Comparative Analysis of K-Means and Fuzzy C-Means Algorithms, International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013.

[8]. Sudipto Guha, Rajeev Rastogi, Kyuseok Shim (1998), CURE: An Efficient Clustering Algorithm For Large

[9]. A. Fahad, N. Alshatri, Z. Tari, A. ALAmri, A. Y. Zomaya, I. Khalil, F. Sebti, and A.Bouras, "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis, "IEEE transactions on emerging topics in computing, 2014.

[10]. T.Garg, Arun Malik, "Survey on Various Enhanced K-Means Algorithms", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 11, November 2014.

[11]. Shi Na, Liu Xumin, Guan Yong, Research on K-means Clustering Algorithm: An Improved K-means Clustering Algorithm, Intelligent Information Technology, and Security Informatics,2010 IEEE Third International Symposium on 2-4 April, (pp. 63-67), 2010.

[12]. Hareesha, K., Shafeeq, A., Dynamic Clustering of Data with Modified K-Means Algorithmî International Conference on Information and Computer Networks, vol. 27,2012

[13]. A.benayed, M.benhalimaand M. alimi, "Survey on clustering methods: Towards fuzzy clustering for Big Data," In Soft Computing and Pattern Recognition (SoCPaR), 6th International Conference of. IEEE, p. 331-336, 2014.

[14]. Libao Zhang, Faming LU, An LIU, Pingping GUO, Cong LIU,îApplication of K-Means Clustering Algorithm for Classification of NBA Guardsî, International Journal of Science and Engineering Applications Volume 5 Issue 1, 2016, ISSN- 2319-7560 (Online).

[15]. A. Sherin, S. Uma, K.Saranya and M. Saranya Vani" Survey On Big Data MiningPlatforms, Algorithms And Challenges".International Journal of Computer Science & Engineering Technology, Vol. 5 No, 2014

[16]. Wen-Jun Zhang, Xiao-Feng Xie, and Zhi-Lian Yang, "Adaptive Particle Swarm Optimization on Individual Level," IEEE, International Conference on Signal Processing (ICSP), Beijing, China, pp. 1215-1218, 2002.

[17]. S. Arora, I. Chana, "A survey of clustering techniques for Big Data analysis," in Confluence the Next Generation Information Technology Summit (Confluence), 5th International Conference. IEEE, p. 59-65, 2014.

[18]. Deepali Virmani, Shweta Taneja, Geetika Malhotra, "Normalization based K means Clustering Algorithm", International Journal of Advanced Engineering Research and Science, Vol-2, Issue-2, ISSN: 2349-6495, Feb. - 2015.

[19]. S. Hari Ganesh, J. James Manoharan, Ph.D., Dr. J.G.R. Sathiaseelan, "Outlier Detection Using Enhanced K-Means Clustering Algorithm and Weight Based Center Approach", IJCSMC, Vol. 5, Issue. 4, pg.453 – 464, April 2016.

[20]. S. Wang, C. Yadav, ET M. Kumar, "Algorithm and approaches to handle large data-A Survey," International Journal of computer science and network, vol 2, issue 3, 2013.