IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

A Survey: Big data Analysis in Healthcare using machine Learning Approach

[1*]A. Thammi Reddy, [2] M. Nagendra

[1]Rayalaseema University, Kurnool, Andhra Pradesh, India
[2]Department of Computer Science, S.K.University, Anantapuram, Andhra Pradesh. India

DOI: https://doi.org/10.26438/ijcse/v7i2.300307 | Available online at: www.ijcseonline.org

Accepted: 10/Feb/2019, Published: 28/Feb/2019

***Abstract*: -**Big Data revolution is transforming the way we live.Healthcare industry generates huge data about every patient but accessing, managing and interpreting the data are critical to creating actionable insights for better care and efficiency.Clinical trends also play a role in the rise of Big Data in Healthcare.Furthermore, big data and machine learning will continually and drastically improve every area of the healthcare industry over the next decade as big data techniques become more refined. This paper presents the role of Big Data analysis in healthcare and various shortcomings of traditional machine learning algorithms.

Keywords: Big Data, Machine learning, Healthcare Analysis, Feature Selection.

I. INTRODUCTION

Healthcare industry generates huge data about every patient but accessing, managing and interpreting the data are critical to creating actionable insights for better care and efficiency.Big data resolutions frequently come with regular innovative data management solutions and vital tackles. Big Data in healthcare is used to predict epidemics, cure disease, enhance the quality of life and avoid unnecessary deaths. Nowadays, healthcare industries conduct their investigation by using big data software that's being held in a cloud. Coming to big data scenario, its three properties, i.e. one have characterized big Data.Volume, 2.Velocity and 3.Variety [4].

Volume refers to the massive amount of data being generated by several sources. Velocity refers to the rate at which this data is being generated and the variety refers to the different type of information being used (e.g. text, images, videos etc.).From the analytical/statistical point of view, Big Data may not be big in volume only; it may be significant in terms of dimensions too. Dimensions, which are often called features in statistics. Big data at times is so massive that the conventional methods of data mining can barely give any information in useful time [5].

Nowadays with so much data all around the world, the trend in healthcare is shifting from cure to prevention. Hospitals and healthcare systems are excellent repositories of big data (like patient records, test reports, medical images etc.) that can be utilized to cut the cost in healthcare, to improve reliability and efficiency, and to provide better cure to patients [7]. But here, we need to deal with the problems of unstructured data and missing values for some fields. Using the different approaches of machine learning and data mining, we can provide better cure to diseases, develop personalized medicines and even prevent the diseases or epidemics [6]. Traditional machine learning algorithms are not suitable to process such a massive amount of data as the time required to Train and analyze data generally exceeds the tolerable time limit. Traditional Machine Learning Algorithms or more specifically traditional machine learning infrastructure works on centralized databases. But in case of big data, the data may be in Petabytes, which is not suitable or sometimes even impossible to store and process on a single machine. So we need to parallelise the traditional approaches and need to either modify the traditional algorithms or come up with some hybrid approaches which can take up the challenge of managing the extensive data set in a distributed environment [4].

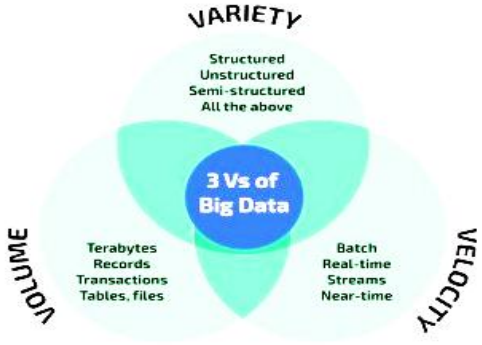

Figure 1.The 3 Vs of Big Data

There are many areas in healthcare where machine learning plays a role already, and the importance of AI in healthcare

will likely increase in the future. The technology will mature, and both legal and cultural hurdles will be surpassed.Broadly speaking, and to give some structure we can identify three areas where ML is being used in healthcare right now:

1. Perception tasks - tasks requiring skills of perception such as vision or hearing etc.
2. Diagnostic assistance
3. Treatment procedures

In recent years, machine learninghas boosted the performance of computers at perception tasks to previously unimaginable levels. It has led to an explosion of use cases in multiple areas, including healthcare. So when working in the domain of healthcare, we need to take care of the tradeoff between the computation time and sub-optimality of the solution. One of the significant challenges of implementing big data machine learning algorithm is to come up with an algorithm which can maintain the perfect balance between the above two trade-offs.

1.1 Big Data Architecture

Big data project management is one of the challenging tasks as we need to collect the data from a different source which have a different format like text, image, audio and video.

Gathering and tracking patient data:

The health industry whips up huge amounts of data in patient records daily. While the digitalization of this data has made it easier to follow, it is not of much use beyond maintaining records. The process itself is tedious and time-consuming. This is where Machine learning comes into the picture. Voice dictation technology helps speed up the documentation process while the actual records are analyzed by Machine Learning to further understand and assist with a patient's diagnosis. Machine Learning performs best when they have a considerable amount of data to investigate to infer predictions. So the abundance of data in the Healthcare sector certainly creates a win-win situation for both. Data nowadays are no better structured so that the old database management system can be utilised for knowledge extraction from the data. In big data, we need to deal with structured, semi-structured and unstructured data [8]. So the first step should be to collect all the data from the relevant sources and then aggregate and store it into one common platform.

Generally, we use open source distribution of Apache Hadoop which provides us with Hadoop Distributed File System storage which takes care of distributed storage and fault tolerance. Once we have the data, we need to process it with lowest computation time in a distributed environment so that we can adopt Map Reduce [9] processing which can take the benefit of HDFS and inbuilt distributed processing to process data as quick as possible. In the processing layer, we generally implement some of the machine learning algorithms which perform some intelligent analysis on the data and supply valuable knowledge which can be usedto generate reports. Thus we can suggest that Big Data architecture can be classified into four layers as shown below in Fig 2.

1.2 Big Data Capability

Data is considered to be the new warehouse for this era. Like the oil when it is unprocessed, it is barely of use. But using the different analytical methods, we can extract the precious information/knowledge hidden in it. Big Data has the potential to make a significant impact on the healthcare industry [7].

Massive data in healthcare domain provide us with the considerable ability to do the predictive analysis of these data and come up with the solution of the new problem which might be derived from the existing resolution of the known issue. Through big data analytics, we can parallelly process the huge volume of data and find out the association among them which can help us to resolve the problem and provide us with the practical solution of the hidden issues. Healthcare domain can effectively use this analytical approach to cut down their cost or be better preparing with the needed equipment or resources, which might be required as per the environment. For example, if there is an outburst of any disease in an area which has earlier occurred in different parts of the world, then using the analytical engine to predict the solution and spread of disease which might happen and thus be better ready with their resources and medicine which might be needed to solve the epidemic problem. Also, hospitals and clinics can use the data to analyze and predict the patient's preferences, which opens a path to possible business. Predictive Analysis where we use various statistical methods, machine learning techniques and, data mining approaches to process, analyze data and predict the outcome for the unknown bag of data. Healthcare domain is still in an early stage to take up the new possibilities, which can be offered by big data solution and use it to do effective decision-making.

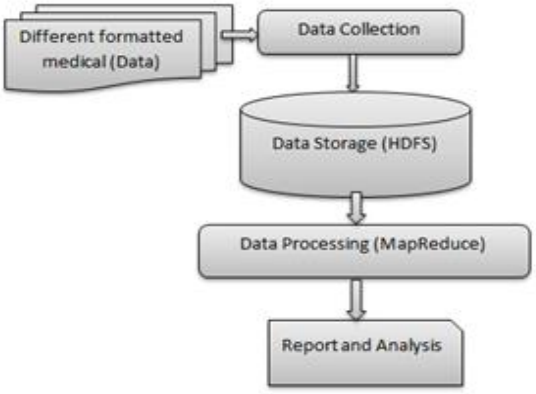

Figure 2. Big data framework

II. RESEARCH CHALLENGES OVER BIG DATA ANALYSIS IN HEALTH CARE FOR CHRONIC DISEASE PREDICTION.

Much of the highly valuable healthcare data is in unstructured or semi-structured form. Added to it, the complex, dynamic and heterogeneous characteristics of the data [1–3] renders it difficult to extract useful information using traditional data analytical tools & techniques.

1. Data are preprocessing stage: Medical data collecting from different sources like sensors, hospitals, labs etc. Missing data is the major problem while in preprocessing.
2. Data Compression: Reducing the volume of data while maintaining essential data such as anatomically relevant data
3. Parallelisation/real-time realisation: Developing scalable/parallel methods and frameworks to speed up the analysis/processing
4. Sharing/security/anonymization: Security issues such as Big Data breaches can be a significant threat in healthcare. Integrity, privacy, and confidentiality of data must be protected.
5. Data integration/mining: Finding dependencies/patterns among multimodal data and the data captured at different time points to increase the accuracy of diagnosis, prediction, and overall performance of the system
6. Validation: Assessing the performance or efficiency of the system

III.MACHINE LEARNING APPROACHES FOR ANALYSIS AND PREDICTION

How is machine learning being used in healthcare?

Machine learning is making massive amounts of data and their analysis available to the medical profession. There have been many exciting developments in this field which is just an extension of artificial intelligence.

Machine learning challenges the traditional, reactive approach to healthcare. It's the exact opposite: predictive, proactive, and preventative—life-saving qualities that make it a critically essential capability in every health system.

Machine learning in medicine has recently made headlines. Google has developed a machine learning algorithm to help identify cancerous tumors on mammograms. Stanford is using a deep learning algorithm to detect skin cancer.

Still, machine learning lends itself to some processes better than others. Algorithms can provide immediate benefit to disciplines with methods that are reproducible or standardised. Also, those with large image datasets, such as radiology, cardiology, and pathology, are strong candidates. Machine learning can be trained to look at images, identify abnormalities, and point to areas that need attention, thus improving the accuracy of all these processes. Long term, machine learning will benefit the family practitioner or internist at the bedside. Machine learning can offer an objective opinion to improve efficiency, reliability, and accuracy.

Data Analysis always has demand in all the industry as it gives the approximate prediction of how the market is growing. Machine learning algorithm helps in making that prediction. To perform prediction, many factors need to be taken into account. We know the class into which the prediction model is going to work. We have mainly focused on supervised or unsupervised learning algorithm covering some of the technique which is mainly used in the healthcare domain. High-dimensional data analysis is a challenge for researchers and engineers in the fields of machine learning and data mining. Feature selection provides an effective way to solve this problem by removing irrelevant and redundant data, which can reduce computation time, improve learning accuracy, and facilitate a better understanding of the learning model or data. For big data, it is always preferable to make some feature selection to reduce the dimensionality of data and then make the prediction [4]. We have summarised various techniques in the following section.

Machine Learning Model in Healthcare Data:

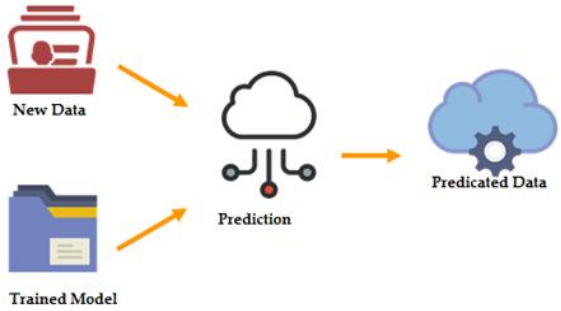

Figure 3. Machine Learning Model in Healthcare data

A. Feature Selection

In the past thirty years, the dimensionality of the data involved in machine learning and data mining tasks has increased explosively. Data with extremely high dimensionality has presented severe challenges to existing learning methods. Feature selection is an essential preprocessing technique used before data mining. In this, we select a subset from the given features set so that we can reduce the computational complexity of the learning algorithm and remove irrelevant/redundant elements to eliminate noise. This step makes the model easier to interpret. Without exhaustive search, no algorithm can guarantee an optimal feature subset, which itself is an NP-Hard problem. So in Big Data problem thorough search cannot be applied. Feature selection, which has been a research topic in methodology and practice for decades, is used in many fields, such as image recognition [10], image retrieval [11], text mining [12], intrusion detection [13], bioinformatics data analysis [14], fault diagnosis [15], and so on.

Components of Feature Selection Algorithm:

Feature selection: is a process where we automatically select those features in our data that contribute most to the prediction variable or output in which we are interested in.Having irrelevant features in our data can decrease the accuracy of many models, especially linear algorithms like linear and logistic regression.

Approaches for Feature Selection:

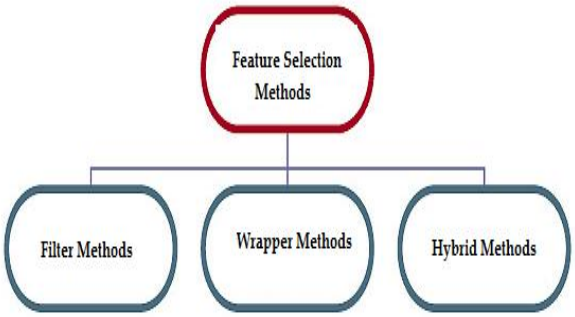

Figure 4. Feature Selection methods

Filter Method

Filter method deals with individual ranking as well as subset selection. The different ranking is based on the evaluation functions such as distance, information, dependence, and consistency excluding the classifier.Because it does not runs the learning algorithm, so it takes less time than the former approach. Hence it is more suitable for the Big Data problem. Correlation-Based feature selection is a well-known example of this type, which is often used for Big Data Problem.

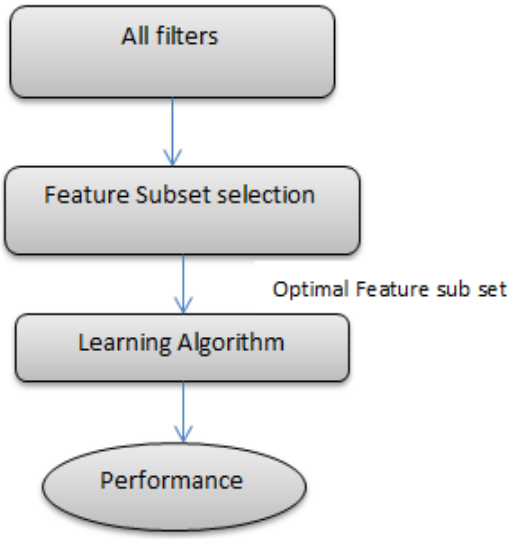

Figure 5. Filter Approach

Advantage and Limitation of Filter approach methods

Advantages:

1. The method is fast and straightforward.
2. It scales well to high dimensional data.
3. It is independent of classifiers.

Disadvantage:

1. The method is generally univariate or low variate.

Wrapper methods

In the wrapper approach, all (Generation Procedure) GP can be taken in combination with the classifier as an evaluation function and generates the relevant feature subset. Wrappers are feedback methods, which incorporate the machine-learning algorithm in the feature selection process, i.e., they rely on the performance of a specific classifier to evaluate the quality of a set of features.

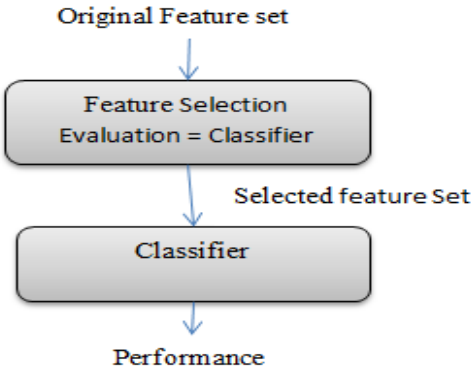

Figure 6. Wrapper methods

Advantages:

The advantage of the wrapper approach is that it selects a near perfect subset and error rate in this method is less as compared to other methods.

Disadvantages:

The major problem of the method is that it is computationally very intensive and it is intended for the particular learning machine on which it has been tested. Therefore, there is a high risk of overfitting than filter techniques.

Hybrid Approaches

It evaluates the features and chooses the best from them to make subsets in further iterations. Different kind of such subsets is compared to taking the learning algorithm into account [11]. Thus, the Hybrid approach takes advantages from both of the above procedures. This approach maintains the trade-off between time and efficiency. So this is more suitable for Big Data about Healthcare then above procedures. But still, there is the scope of the sub-optimal solution as the two best features are not necessarily the best two [16].

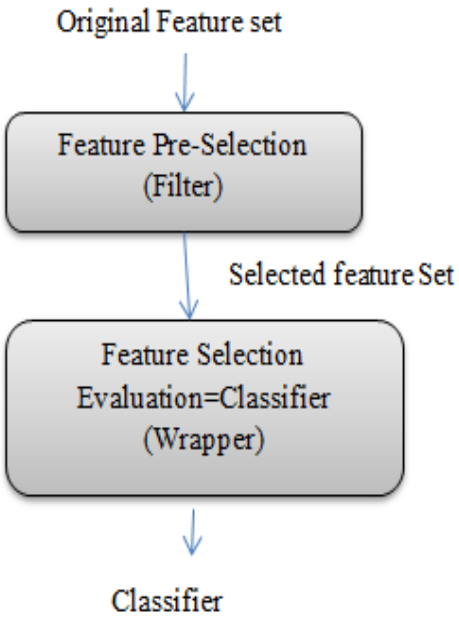

Figure 7. Hybrid Approaches

Advantage:it manages to improve the efficiency and prediction accuracy with the better computational cost for high dimensional data. Three benefits of performing feature selection before modelling our data are:

- **Reduces Overfitting**: Less redundant data means less opportunity to make decisions based on noise.
- **Improves Accuracy**: Less misleading data means modelling accuracy improves.
- **Reduces Training Time**: It reduces the complexity of a model and makes it easier to interpret.

IV. CLASSIFICATION

Classification is a supervised machine learning technique, where the classification model is built based on the input data and mapped target class provided to the classifier [17]. The traditional machine learning classification algorithms take as input, a set of data where target class is already known and builds a classifier model for the target class based on features of input data. After that test data set is provided to the classifier to check whether the model is mapping to correct target class for the test data set whose level was unknown to the system. In the present day, with the growth of big data, traditional classification algorithm has some limitation as it was meant for the homogenous domain and task and sometimes it has highcomputational cost for a large volume of data [4]. We list downthe various traditional classification techniques which aremainly used in the healthcare domain and their shortcomingsfor the current age of big data.

Decision Tree (DT):

Decision Tree is a predictive model of classification, which can be viewed as a Tree like structure [14]. It is simple and gives a fast and accurate result based on the examples, which has been used in building it. With the rise of massive data set, big data, there are some drawbacks of the traditional decision tree algorithm as building the decision tree with such a huge amount of data might be time-consuming and if we divide the data into small partition then the challenge is to keep the information needed at the time of computation local to the nodes so that communication cost can be reduced. In the healthcare domain, the decision tree is one of the most widely used classification approaches as it can simply classify the patient based on question and assign the patient an appropriate class label. Below we give a generalised decision tree as shown in Fig 8.

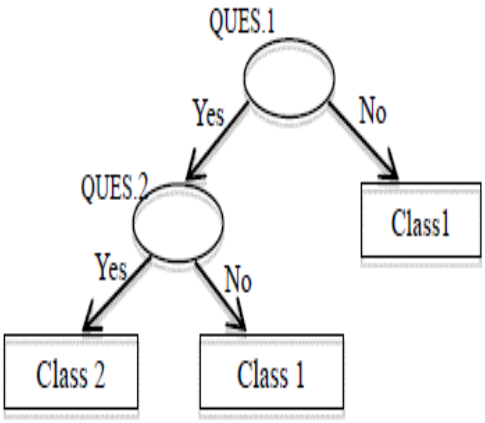

Figure 8. Decision Tree

Support Vector Machine (SVM):

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is the number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyperplane that differentiates the two classes very well (look at the below snapshot). Among the entire traditional machine learning approaches, support vector machine is the latest and most popular, but its performance degrades when the data set has a large amount of noisy data.SVM takes time in the training of the classifier and then the prediction is very much faster [18]. SVM classifies the data point based on the support vector and hyperplane, which separates the data points in higher dimensional space. SVM is generally used in combination with other technique to remove the problem of noisy data **Fig. 9**which is depicted below shows how SVM classify the data by nonlinearly mapping it into a new dimension and then having a hyperplane to draw the decision boundary.

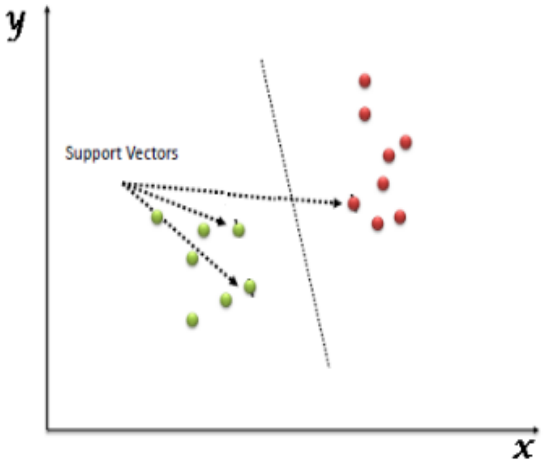

Figure 9 : Support vector machine Classification

Neural Network (NN):
Neural Network is one of the other machine learning algorithms which showed a lot of modification. Neural Network is an adaptive learning model which adjusts the weight of the connecting links between its neuron. Among all artificial neural networks, Multilayer Perceptron (MLP) is one of the widely used neural-network Neural Network models gives an accurate result but the problem of computation time in case of big data needs to be tackled by hybridisation of it with other model or using it at the later stage

K-Nearest Neighbor (K-NN):
K-Nearest Neighbour model of classification is one of the most straightforward classification algorithms which work on the classifying the data set based on the nearest neighbour of the existing class label of the already trained model [24]. It takes into account 'k' neighbour to find out the conclusion for the data class. Generally, some optimisation algorithm is applied before the k-nearest neighbour to reduce the feature set as in the case of big data it would reduce up the computational cost. It gives a better result than the Bayesian method in case of text mining [24].

Bayesian Methods
Bayesian Method is based on the Bayes theorem. Naïve Bayesian Classifier has very good accuracy in classification for a large set of data [25]. The problem with this method is that it takes the entire attribute independently; thus if the characteristics are independent, then only Naïve Bayesian classifier gives it full accuracy [23]. Naive Bayesian classifier is a robust statistical classifier when it comes to efficiency.But as the name suggests it is naive and takes the assumption that all attributes are independent of each other. So if we can find independent characteristics either by preprocessing(e.g. feature selection)or without preprocessing this classifier can perform well in Healthcare Domain.

Random Forest
Random forests can be used to estimate feature importance's, and to rank which features have the best predictive power.

- The same random forest algorithm or the random forest classifier can use for both classification and the regression task.
- Random forest classifier will handle the missing values.
- When we have more trees in the forest, random forest classifier won't overfit the model.
- Can model the random forest classifier for absolute values also.

V. CLUSTERING

Clustering is used for analysing and grouping data which does not include pre-labelled class or even a class attribute at all. Clustering algorithm makes the groups or clusters of homogenous data. It is an unsupervised learning technique. Unsupervised because opposed to classification it is not provided with any class labels initially. It makes the classes/clusters based on the similarity between data points, following the rule that data points within the same cluster should have higher similarity and data points between different clusters should have the least resemblance. The similarity measures being used are often Euclidian distance, Cosine Measure, Pearson correlation, Jaccard Measure etc. Clustering needs no prior information about the data to work upon. So it can be used for microarray data in which we have very little details about the genes. Tapia et al. analyzed the gene expression data with the help of a new hierarchical clustering approach using a genetic algorithm. This approach is beneficial concerning Big Data because it can represent the information in compact form by creating suitable clusters with almost no loss of information [21].

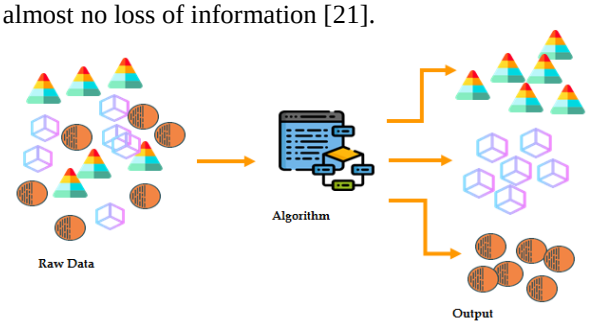

Figure 10. Clustering Process

From the above figure given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. Many thousands of clustering algorithms exist based on more than a dozen different approaches. Traditionally, we have two basic methods - partition and hierarchical clustering algorithms

Partitioned Clustering

Partition methods represent one-level clustering when as a result we obtain a set of clusters and no-hierarchy. While the given cluster is not being analyzed, it is stable or consists of the components in turn. This algorithm presents the number of clustersbeforehand.In other words, the whole data set is partitionedinto a predefined number of partitions, such that none of suchbarriers is empty and every data point belongs to one clusteronly. The choice of cluster centroid and similarity measurefurther divides this approach into two categories, i.e. K-Meansand K-Medoids. K-Means is more popular than KMedoids.

In K-Means the distance or the similarity measure iscalculated between the centroid of the cluster and the datapoint. The data points in the same group should have highersimilarity, and in different clusters, they should have the leastresemblance [21]. Whereas in K-Medoids the medoids of theclusters are used in place of the centroid. A medoid is a centrepoint of a cluster which exists in the data set. Belciuget al. The methodology presented in [22] used the clusteringtechnique to detect recurrence of breast cancer.

Hierarchical Clustering

Divisive hierarchical algorithms continue to split clusters regardless of their atomicity. i.e. Agglomerative hierarchical algorithms have the same weakness - they can unify non-similar clusters during the work. In Hierarchical Clustering [21] we don't need to define thenumber of clusters in advance. Based on it working approachit can be divided into two categories

Hierarchical clustering algorithms are either top-down or bottom-up.

1. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering is therefore called hierarchical agglomerative clusteringHAC.

2. Top-down clustering requires a method for splitting a cluster. It proceeds by splitting clusters recursively until individual documents are reached.

HACProcess:

- Assign each object to a separate cluster.
- Evaluate all pair-wise distances between clusters.
- Construct a distance matrix using the distance values.
- Look for the pair of clusters with the shortest distance.
- Remove the pair from the matrix and merge them.
- Evaluate all distances from this new cluster to all other clusters, and update the matrix.
- Repeat until the distance matrix is reduced to a single element.

From both of the above approach the stopping criterion isusually the number of clusters to be achieved; once therequired amount is reached the algorithm can be stopped

VI. COMBINATION OF DIFFERENT CLASSIFIERS

There is no single traditional data mining technique whichcan perform well for big data in the healthcare domain, so it isgenerally best to fuse the different approaches and dothe classification using that classifier [18]. Association is one of the favourite techniques for getting the relationship betweendata, and it provides us with the relationship between the data. There are many areas in the healthcare domain when it is useful tounderstand the connection between diseases so that similar treatmentcan be followed. In association, the Apriori algorithm which getsthe relationship between items and separates the frequentitems and infrequent items. Some of the related work hasshown that the fusion of different technique gives a better result.Support vector machine is generally combined with PSOoptimization techniques, which gives optimizes the inputfeature set first, and then the classifier is applied on that toseparate the data, or vice versa can be applied. Similarly, KNN is a user with Fuzzy logic first to decrease thecomputation time [20]. Thus, is an excellent way to go forhybridisation and get the result?

VII. CONCLUSION

Our survey concludes thatit has shown the role of Big Data analysis in Health care for chronic disease prediction. Furthermore, we also explained how machine learning is being used in healthcare.This paper has clearly shown that traditional machine learning algorithms are complicated when it processes the unstructured or semi-structured data lack of scalability, time-consuming on computation for large in size data. Sowe recommend that it's essential to fine-tune or transform the existingalgorithm to make it appropriate for the big data problem. Further study of the issue is still required; there is a tremendous growth in this area and a new dimensionas now we can extract more valuable information and hiddenknowledge, which can providean innovation in the medicalcases.

REFERENCE

[1] F.J. Martin-Sanchez, V. Aguiar-Pulido, G.H. Lopez-Campos, N. Peek, L. Sacchi, Secondary use and analysis of big data collected for patient care, IMIA Yearbook 26 (2017) 1–10, http://dx.doi.org/10.15265/IY-2017-008.

[2] Y. Wang, N. Hajli, Exploring the path to big data analytics success in healthcare, J.Bus. Res. 70 (2017) 287–299, http://dx.doi.org/10.1016/j.jbusres.2016.08.002.

[3] B. Cyganek, M. Graña, B. Krawczyk, A. Kasprzak, P. Porwik, K. Walkowiak, M. Woźniak, A survey of big data issues in electronic health record analysis, Appl.Artif. Intell. 30 (2016) 497–520, http://dx.doi.org/10.1080/08839514.2016.1193714.
[4] Suthaharan, Shan. "Big data classification: Problems and challenges in network intrusion prediction with machine learning." ACM SIGMETRICS Performance Evaluation Review 41.4 (2014): 70-73.
[5] RAO, BLS PRAKASA. "of the Notes: Brief Notes on BIG DATA: A Cursory Look." (2015).
[6] Raghupathi, Wullianallur, and Viju Raghupathi. "Big data analytics in healthcare: promise and potential." Health Information Science and Systems 2.1 (2014): 3.
[7] Hermon, Rebecca, and Patricia AH Williams. "Big data in healthcare: What is it used for?." (2014).
[8] Groves, Peter, et al. "The 'big data revolution in healthcare." McKinsey Quarterly (2013).
[9] Rama Satish, K. V., and N. P. Kavya. "Big data processing with harnessing Hadoop-MapReduce for optimising analytical workloads."Contemporary Computing and Informatics (IC3I), 2014 International Conference on. IEEE, 2014.
[10] Yu, Lei, and Huan Liu. "Feature selection for high-dimensional data: A fast correlation-based filter solution." ICML. Vol. 3. 2003.
[11] Liu, Huan, and Lei Yu. "Toward integrating feature selection algorithms for classification and clustering." Knowledge and Data Engineering, IEEE Transactions on 17.4 (2005): 491-502.
[12] Liu, Huan, and Hiroshi Motoda, eds. Computational methods of feature selection. CRC Press, 2007.
13] Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." International Journal of Bio-Science and Bio-Technology 5.5 (2013): 241-266.
[14] Wang, Dingxian, Xiao Liu, and Mengdi Wang. "A DT-SVM Strategy for Stock Futures Prediction with Big Data." Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on. IEEE,2013.
[15]. Fahad, Adil, et al. "A survey of clustering algorithms for big data: Taxonomy and empirical analysis." Emerging Topics in Computing, IEEE Transactions on 2.3 (2014): 267-279.
[16]Dharavath, Ramesh, and Abhishek Kumar Singh. "Entity Resolution-Based Jaccard Similarity Coefficient for Heterogeneous Distributed Databases." Proceedings of the Second International Conference on Computer and Communication Technologies. Springer India, 2016.
[17]Kesavaraj, G., and S. Sukumaran. "A study on classification techniques in data mining." Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on.IEEE, 2013.
[18] Cavallaro, Gabriele, et al. "On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods." (2015).
[19]Jen, Chih-Hung, et al. "Application of classification techniques on development an early-warning system for chronic illnesses." Expert Systems with Applications 39.10 (2012): 8852-8858.
[20]Zuo, Wan-Li, et al. "Effective detection of Parkinson's disease using an adaptive fuzzy k-nearest neighbour approach." Biomedical SignalProcessing and Control 8.4 (2013): 364-373.
[21] Fahad, Adil, et al. "A survey of clustering algorithms for big data: Taxonomy and empirical analysis." Emerging Topics in Computing, IEEE Transactions on 2.3 (2014): 267-279.
[22]Belciug, Smaranda, et al. "Clustering-based approach for detectingbreast cancer recurrence." Intelligent Systems Design and Applications(ISDA), 2010 10th International Conference on. IEEE, 2010.
[23] Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." International Journal of Bio-Science andBio-Technology 5.5 (2013): 241-266.
[24] Bijalwan, Vishwanath, et al. "KNN based machine learning approach for text and document mining." International Journal of Database Theory and Application 7.1 (2014): 61-70.
[25] Gelman, Andrew, et al. Bayesian data analysis. Vol. 2. London:Chapman & Hall/CRC, 2014.

