International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

A Review on Big Data Mining in Healthcare

Chaitali S. Suratkar[1]

Assiststant Prof, Dept. of IT, YCCE College, Nagpur.

Abstract - *Health Informatics is a combination of information science and computer science within the realm of healthcare. The amount of data generated by the healthcare industry is becoming tough to manage and to examine it in efficient manner for future use. The analysis of patient's data is becoming more important, to evaluate the medical condition of patient and to prevent and take precautions for future. Managing the huge volume of data has many problems interrelated to data security, data integrity and inconsistency. To obtain the best services and care for the patients, healthcare organizations in many countries have proposed various models of healthcare information systems. These models for personalized, predictive, participatory and preventive medicine are based on using of electronic health records (EHRs) and huge amounts of complex biomedical data and high-quality – omics data. Data mining and Big Data analytics are helping to realize the goals of diagnosing, treating, helping, and healing all patients in need of healthcare, with the end goal of this domain being improved Health Care Output. In this paper, the various applications of big data mining techniques have been analyzed to improve the healthcare systems.*

***Key Words*: Data Mining, data Mining in Healthcare, Health Informatics**

1. INTRODUCTION

Big data in healthcare and medicine refers to these various large and complex data which they are difficult to analyze and manage with traditional software or hardware. Big data analytics covers integration of heterogeneous data, data quality control, analysis, modeling, interpretation and validation. Application of big data analytics[1] provides comprehensive knowledge discovering from the available huge amount of data. Health Informatics (as in all its subfields) can range from data acquisition, retrieval, storage, analytics employing data mining techniques, and so on. Big Data can be define by five V's: Volume, Velocity, Variety, Veracity, and Value. Volume pertains to vast amounts of data, Velocity applies to the high pace at which new data is generated, Variety pertains to the level of complexity of the data, Veracity measures the genuineness of the data, and Value evaluates how good the quality of the data is in reference to the intended results. Big data mining can aid in analyzing medical operation indicators of hospitals for a period to help hospital administrators provide data support for medical decision-making. In this manuscript, the various algorithms of big data mining techniques have been analyzed to improve the healthcare systems.

The term big data is described by the following characteristics:

value, *volume*, *velocity*, *variety*, *veracity* and *variability*, denoted as **6 "Vs"** The *volume* of health and medical data is expected to raise intensely in the years ahead, usually measured in terabytes, petabytes even yottabytes. Volume refers to the amount of data, while *velocity* refers to data in motion as well as and to the speed and frequency of data creation, processing and analysis. Complexity and heterogeneity of multiple datasets, which can be structured, semi-structured and unstructured, refer to the *variety*. *Veracity* referrers to the data quality, relevance, uncertainty, reliability and predictive value while *variability* regards about consistency of the data over time. The *value* of the big data refers to their coherent analysis, which should be valuable to the patients and clinicians.

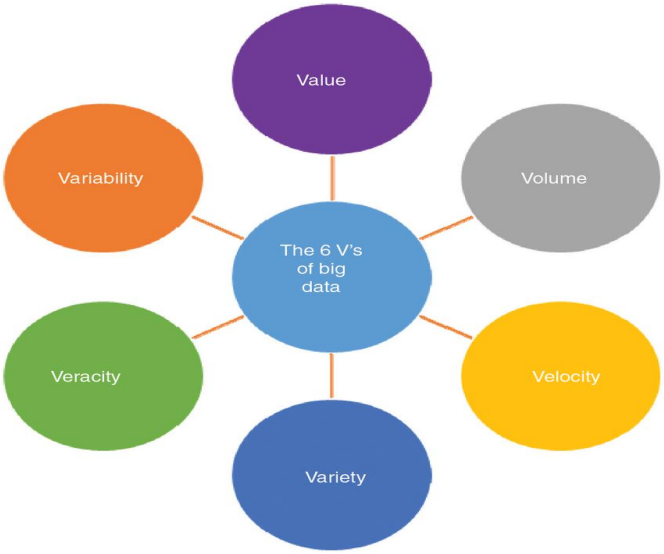

Fig -1: Charecteristics of big data on 6V's

Medical big data can be used to improve healthcare quality, predict epidemics, increasing analytical abilities, cure disease, build better health profiles, improve quality of life, improving outcome, avoid preventable deaths, build better predictive models and reducing resource wastage . Big data can be used in understanding the biology of a disease by integrating the available large volumes of data to build meaningful relational models. Medicinal big data drives the progressions behind models of treatment. So with this kind of technology, we can understand so much about a patient, information as early in their life as possible, collecting warning signs of serious diseases at an early stage for a faster and cheaper treatment. Big data analysis in the medical field will ensure that even the smallest of detail will

be taken into consideration. Following are the top benefits of big data in healthcare field: Facility performance optimization, Energy cost reduction, Ease of accessibility of information, Real time update and alerts. Proactive maintenance of equipment's, Reducing the healthcare delivery cost, Reducing the costs of research and growth.

2. APPLICATION OF BIG DATAMINING IN MEDICAL FIELD

In the field of business and marketing, the application of data mining has been implemented and may be ahead of healthcare. But this is not the case now. Effective mining applications have been actualized in the medical field, some of which are depicted beneath. With the assistance of medical big data and robust mining methods and model building solution, we can identify patients with high risk health condition. This information can be bridled by doctors and medical staffs to identify the condition, so they can take steps to improve quality of healthcare and to prevent health problems in the future . For example, Cancer is a serious illness which can be prevented and cured with the help of big data analytics. Cancer is quickly devastating individuals over the world. Big data can battle disease all the more viably. Healthcare suppliers will have upgraded capacity to recognize and analyze infections in their beginning periods, relegating more adequate treatments in view of a patient's genetic makeup, and direct medication measurements to limit symptoms and enhance viability. It will likewise fantastic help to parallelization and help in mapping the 3 billion DNA base sets.

The reliance of health care on data is increasing. Medical researchers, physicians, and health care providers face the problem to use stored data efficiently when more medical information systems with large database are used. The medical information system databases contain many data such as patient records, physician diagnosis, and monitoring information where the data has been useful in many medical decision support systems to save lives. A medical decision support systems are systems that help in the decision making process in the medical domains such Clinical Decision Support Systems (CDSS), medical imaging, and Bioinformatics . The contributions of these systems are to reduce medical errors and costs, earlier disease detection, and to achieve preventive medicine. The advantages of using computerized Clinical Decision Support Systems(CDSS) are the decision support systems can help to manage overloaded data and turn them into knowledge, reduce the complexity of the work such as automatic complex workflows, and help to identify reducing the errors, time, and variety of practices . Continuous usage of the information systems result to the size of the database increasing. Therefore the usage of knowledge discovery and data mining in the database (KDD) for the growing databases is important. knowledge discovery(KDD) attempts to gather knowledge by identifying relations from the data sets to help predictions. knowledge discovery (KDD) utilization is increasing in medical informatics and researchers have used it in many areas such as statistics, machine learning, intelligent databases, data visualization, pattern recognition, and high performance computing .

3. DATA MINING ALGORITHMS IN HEALTHCARE

Healthcare covers a detailed processes of the diagnosis, treatment and prevention of disease, injury and other physical and mental impairments in humans . The healthcare industry in most countries are evolving at a rapid pace. The healthcare industry can be regarded as place with rich data as they generate massive amounts of data including electronic medical records, administrative reports and other benchmarking finding . These healthcare data are however being under-utilized. Data mining is able to search for new and valuable information from these large volumes of data. Data mining in healthcare are being used mainly for predicting various diseases as well as in assisting for diagnosis for the doctors in making their clinical decision. The discussion on the various methods used in the healthcare industry are discussed as follows.

3.1 Anomaly Detection

Anomaly detection is used in discovering the most significant changes in the data set. Anomaly detection is a technique used to identify unusual patterns that do not conform to expected behavior, called outliers. It has many applications in business, from intrusion detection (identifying strange patterns in network traffic that could signal a hack) to system health monitoring (spotting a malignant tumor in an MRI scan. Three different anomaly detection method, standard support vector data description, density induced support vector data description and Gaussian mixture to evaluate the accuracy of the anomaly detection on uncertain dataset of liver disorder dataset which is obtained from UCI. The method is evaluated using the AUC accuracy.

3.2. Clustering

Clustering is the method of converting a group of abstract objects into classes of similar objects. Clustering is a method of partitioning a set of data or objects into a set of significant subclasses called clusters. It helps users to understand the structure or natural grouping in a data set and used either as a stand-alone instrument to get a better insight into data distribution or as a pre-processing step for other algorithms. The clustering is a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data . The algorithms used in the vector quantization method are k-means, k-mediods and x-means.

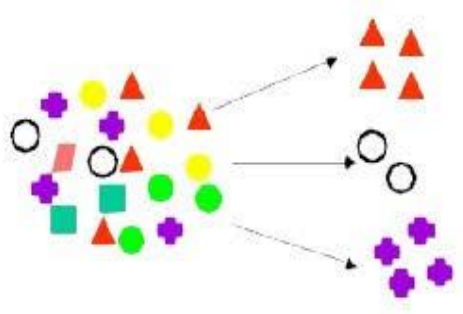

Fig -2: Example of Clustering

3.3. Classification

Classification is the discovery of a predictive learning function that classifies a data item into one of several predefined classes .It is a Data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known. For example**:** If the patients are grouped on the basis of their known medical data and treatment outcome, then it is considered as classification.

3.4 Statistical

The MTS algorithm is being extensively applied in multivariable statistical analysis. The Mahalanobis distance (MD) is used to build statistical judgements to distinguish one group from another and the Mahalanobis space (MS) is used to represent the degree of abnormality of observations from the known reference group. The class imbalance problems are very much prevalent in the healthcare datasets. Usage of the data mining algorithms are often affected with skewed distribution when using skewed or imbalanced data sets. This problem often leads to the tendency of producting highly predictive classification accuracy over the majority class and poor accuracy over the minority class. Having such a nature to distinguish the degree of abnormality of observations, this method would be a good method to test on the real data set pressure ulcers.

3.5. Discriminant Analysis

Linear discriminant analysis (LDA) is widely used in discriminant analysis to predict the class based on a given set of measurements on new unlabeled observations. The linear discriminant analysis is the conditional probability density function of the predictors follows a normal distribution based on the given class value. The algorithm's ability to capture statistical dependencies among the predictor variables indicates that this algorithm would be suitable to explore the linear constraint of this study to discovery the synergy between motor and non motor symptoms.

3.6 Decision Trees

J48 algorithm: this algorithm's name is derived from its tree-like structure and is based on supervised learning techniques. It is a frequently used algorithm due to its ease of implementation, low cost, and reliability. Decision trees' roots consist of decision nodes, branches, and leaves . In the WEKA software, the J48 algorithm uses the rules of the C4.5 algorithm. Therefore, in WEKA, the J48 algorithm is considered a C4.5 algorithm. The C4.5 algorithm can manage numerical values, large data quantities, and datasets with missing values. The C4.5 algorithm uses a threshold value to divide the data into two ranges. The threshold value is selected to provide the most information from the raw data and is determined by sorting the attributes and selecting the average value of the attributes.

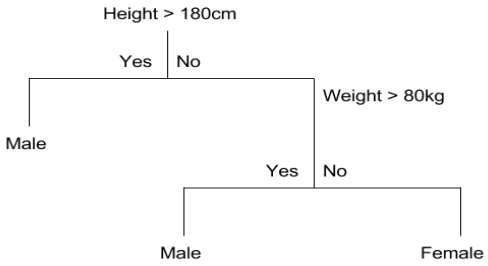

Fig -3: Classification of data by decision tree

3.7 K-Nearest Neighbor

k-nearest neighbors algorithm (*k*-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression: In *k-NN classification*, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor

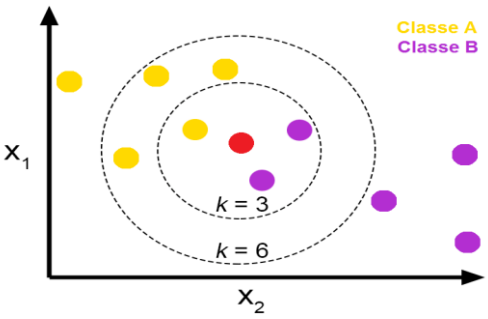

Fig -4: Classification of data by K-Nearest Neighbor

The k-nearest neighbor is an instance based classifier method. The parameter units consists of samples that are used in the method and this algorithm then assumes that all instances relate to the points in the *n*-dimensional space.This algorithm would be suitable if the training data set is large as this algorithm is very time consuming when each of the sample in training set is processed while classifying a new data and this process requires a longer classification time.

3.8 Logistic Regression

Logistic regression: logistic regression measures the relationship between a response variable and independent variables, like linear regression, and belongs to the family of exponential classifiers. Logistic regression classifies an observation into one of two classes, and this algorithm analysis can be used when the variables are nominal or binary. The data are analyzed after the discretization process for the continuous variables, similar to the Bayesian group. Logistic regression is a machine learning algorithm for classification. In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function

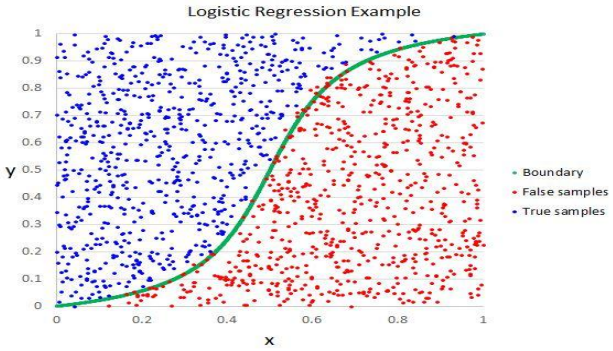

Fig -5: Classification Of data by Logistic Regression

Logistic regression (LR) is a method that would use the given set of features either continues, discrete, or a mixture of both types and the binary target, the LR then computes a linear combination of the inputs and passes through the logistic function [2]. This method is commonly used because it is easy to implementation and it provides competitive results. The LR works well for larger datasets. Logistic regression is designed for this purpose (classification), and is most useful for understanding the influence of several independent variables on a single outcome variable

3.9 Bayesian Classifier

Naive Bayes: the Naive Bayes algorithm is based on the Bayesian theorem and operates on conditional probability. Despite its simplicity, it is a powerful algorithm for predictive modeling. Additionally, the Naive Bayes classifier works quite well concerning real-world situations. An example is spam filtering, which is a well-known problem for which the Naive Bayes classifier is suitable. As with the BayesNet[2] algorithm, there should be no missing data in this algorithm and the variables must be discrete.

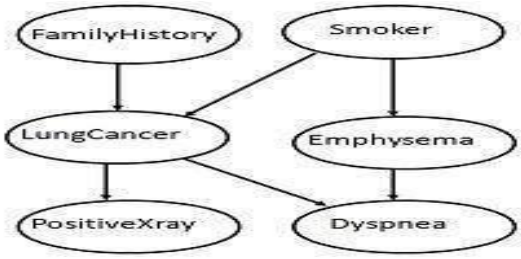

Fig -6: Classification of data by Bayesian Classifier

The Bayesian classifiers is well known for its computational efficient and ability to handle missing data naturally and efficiently. This method would be a good approach if there data sets are suffering from missing data.

3.10 Support Vector

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is 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 hyper-plane that differentiates the two classes very well.

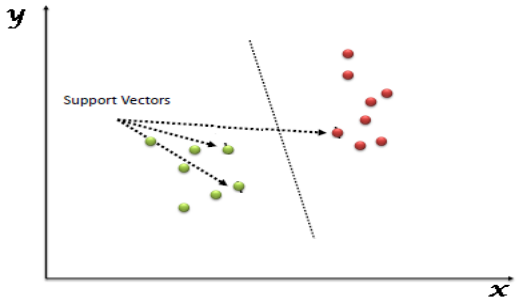

Fig -7: Example of classification by . Support Vector

Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line). The support vector method (SVM) is to be advantageous in handling classification tasks with excellent generalization performance. The method seeks to minimize the upper bound of the generalization error based on the structural risk minimization principle. The SVM training is equivalent to solve a linear constrained quadratic programming problem. The method is very commonly used in medical diagnosis. The training error rate can be controlled by changing the features in the classifiers

Conclusion

The data mining has played in an important role in healthcare industry, especially in predicting various types of diseases. The diagnosis is widely being used in predicting diseases, they are extensively used in medical diagnosing. In conclusion, there is different algorithm can be used for data mining method to solve the issues in the healthcare data sets. In order to obtain the highest accuracy among classifiers which is important in medical diagnosing with the characteristics of data.

REFERENCES

[1] Laura Elezabeth, Ved P. Mishra, Joanita Dsouza."The Role of Big Datab Mining in healthcare Application"International Conference on Reliability, Infocom Technologies and Optimization, August 29-31,2018Amity University.

[2] Neesha Jothia, Nur'Aini Abdul Rashidb, Wahidah Husain," Data Mining in Healthcare – A Review" Procedia Computer Science 72 (2015) 306 – 313

[3]

[4] Dimple*,"A Review on Data Mining Techniques used in healthcare industry" International Journal in Multidisciplinary and Academic research, Vol ,No 1 Feb-March 2014(ISSN 2278-5973)