Big Data for Health Care Analytics using Extreme Machine Learning Based on Map Reduce

Sivakumar Karuppan, N. S. Nithya, Revathy Ondimuthu

Abstract – A large volume of datasets is available in various fields that are stored to be somewhere which is called big data. Big Data healthcare has clinical data set of every patient records in huge amount and they are maintained by Electronic Health Records (EHR). More than 80 % of clinical data is the unstructured format and reposit in hundreds of forms. The challenges and demand for data storage, analysis is to handling large datasets in terms of efficiency and scalability. Hadoop Map reduces framework uses big data to store and operate any kinds of data speedily. It is not solely meant for storage system however conjointly a platform for information storage moreover as processing. It is scalable and fault-tolerant to the systems. Also, the prediction of the data sets is handled by machine learning algorithm. This work focuses on the Extreme Machine Learning algorithm (ELM) that can utilize the optimized way of finding a solution to find disease risk prediction by combining ELM with Cuckoo Search optimization-based Support Vector Machine (CS-SVM). The proposed work also considers the scalability and accuracy of big data models, thus the proposed algorithm greatly achieves the computing work and got good results in performance of both veracity and efficiency.

Keywords -- Map reduce, Machine Learning, Big Data Analytics, EHR, CS-SVM.

I. INTRODUCTION

As the technology is evolving in enormous kind also the data size is extended correspondingly. People are living in the world of data. Data sets are in the arrangement of large volume of storage spaces possibly in petabytes too. Looking Big Data in Medical field involves a variety of data of structured, semi-structured and unstructured form. Raw data are available in the form of complex reports, patient's medical history, and electronics test results. These medical reports are in the form of structured and unstructured data. There is no problem to use structured data for risk prediction model. But, there is a lot of valuable information buried in unstructured data format because this data is very discrete, complex, multidimensional and noisy In Health care Electronic Health Records (EHS) which consists of patient's disease records that uses better clinical decision making. This data in clinical healthcare provide the way to perform predictive analysis. Big data analytics has a great potential to process large amount of data in parallel and solution of hidden problems can also be found. This analytical approach is useful for the low expense of execution time on the huge amount of data sets. For example, any disease that has occurred earlier in many parts of the world, prediction of that disease can be done efficiently.

Revised Manuscript Received on February 15, 2020.

Sivakumar Karuppan, Assistant Professor, Department of CSE, JCT College of Engineering and Technology, Coimbatore, India.

Dr. N. S. Nithya, Associate Professor, Department of CSE, KSR College of Engineering , Tiruchengode, India.

Revathy Ondimuthu, Assistant Professor, Department of ECE, Hindustan College of Engineering , and Technology, Coimbatore, India.

Although healthcare data will reconstruct the information to investigate and predict the patient preferences. In predictive analysis, the statistical methods, data mining, and machine learning methods are employed to analyze, process, and predict the features for undiscovered data. Healthcare domain has a lot of possibilities to provide better cure for disease using different analytical tools. The ML algorithms [2] such as Naive Bayesian (NB), Decision Tree(DT), KNN, and Neural Network (NN) are used to handle the structured data. The problem in health care data is processing of unstructured data. Many researches focused on unstructured data processing facing issues regarding training of large data sets in unsupervised learning [3]. Hadoop Map Reduce framework provides the platform for data distribution. It is based on the master/slave architecture. The advantages of using map reduce is that it provides scalable computations within one iteration. Some algorithms want the futuristic information of one node to another at the time of computation. Probabilistic graphical models (PGMs) are a scrutinized machine learning framework for looking these kinds of problem structures. The proposed approach is measured in terms of accuracy and prediction of number of diseased cases from dataset.In many of the cases, the Apache Hadoop framework is used as it is available open source and produces HDFS which gives way for the distributed storage and can be tolerant towards fault. Map reduce is the widely used programming framework for Hadoop that can made used for the processing of large data in a quick manner. The dataset here are segmented into two halves the former is the training data and the latter is the testing data. The ML algorithms are also implemented for the making the system to behave intelligently on the data and can also produce vital information which are used to produce standard reports in the processing layer. In-order to diagnose the disease, the ML approaches are mostly used. These Machine Learning technique helps in arriving at a better decision on the plans of the treatment and also helps providing a better health-care solutions. In clinical database the large amount of data such as the patients' data , images that are related with radiographies, medical reports and other sensor data. These information overload will obviously cause the size and complexity of the database to increase.

Speaking about the utitlities of distributed systems such as Hadoop and Mapreduce is more advntage in clinical healthcare research ares because of its large storage space and computation for huge set of data. Extreme Machine Learning is made used to bring solution to many clinical domain since it has the capacity to bring meaningful information from the samples that are trained. . The remaining work is arranged as follows: The section II of the paper gives information on the available literature.

Section III discusses on the analytics in the Health Care. Section IV elucidates the proposed work . The paper is concluded in Section V.

II. LITERATURE SURVEY

A lot of research has been carried out in big data throughout recent years. Muhammad U. S *et al* presented the survey about big data analytics [1] in medical field. The common ML model such as Neural Network depicted lot of difference. In NN, the process is based on neurons that are link together by fine-correcting the weight to reach an optimal solution. NN is comfortable for solving problems that is often mentioned as Multilayer Perception (MLP). The working principle of such system is resembled to human brain. First, NN is trained to perform classification. After accomplishment of training, testing is performed on input data. The drawback of the system is enforcing the neural network alone is computationally exhaustive. J. Han [3] presented the ML techniques by using K means approach. In the focus of extraction of data from the databases by making use of the clustering, the acquired data that are raw must be pre-processed that includes, correction of data that are not relevant, the selection of attributes and the transformations. The process of data-cleaning has to be performed as the raw data are prone to have missing values , some outliers and also irrelevant data. According to the concept of clustering , some of the attributes may be not appropriate and hence these have to be eliminated. The Normalization is needed on the attributes that are selected to make it acceptable by the algorithms considered. The patterns that evolve from the clusters are again evaluated using performance measures and only the patters that are interesting may turn out as knowledge. In traditional data sets, RDBMS is used for handling bulky dataset propose work focused the big data that are stored in data Architectures like High Performance Computing System (HPCS) to store and analyse giant data sets reliably. Our work is motivated by these previous works and complements them in many ways.

III. BIG HEALTH CARE DATA FOR ANALYTICS

Most commonly, The data from the health care domain is comprised of many large and complex information pertaining to the patient. The recent advancement in medical domain also provide sensor data captured through many electronic devices and mobile devices which further increases the size and complexity of the data. More details about the research are going in a certain way by collecting the information regarding the clinical data from every patient from various hospitals. Finally, the records of details like clinical pictures, CT-scan reports, MRI- scan reports, lab records, surgical records, insurance information are periodically included into the clinical databases. For attaining proper output the record used in hospitals are used such as clinical pharmacy images, Electronic data such as X-Rays images, the images from MRI scans, the post surgery reports , the information pertaining to medication and other general data such as insurance and various other information that are related to a patientare endlessly being enclosed into aid databases.Hence , the dimensions of healthcare database are increasing enormously. As medical health care are larger and they are split up into training data sets and test data for decision making paradigm. Data used for training which contains maximum volume of EHR data's (nearly around eighty percentage from the data set). Training data set are used to perform the prediction actions and also to regulate the input features on the learning machine network. The test data which have the input data only to test the final output to validate regarding actual predictive nature of the algorithm. The test data holds twenty percentage of the clinical data set.

A. Risk Factor Prediction

Normally the risk factor is identified by machine learning and deep learning algorithm. This work focuses on the accuracy measures in point, we utilize the ELM algorithm with CS-SVM [4] for better optimization and accuracy. For Structured data, Naïve Bayes (NB), CS-SVM, Decision tree (DT) Machine learning algorithm is employed to find the risk of fatal disease. For unstructured data we improved the ELM into Optimized Extreme Learning Algorithm to predict the fatal disease base on the training data sets.

For structured data, the prediction using classification approaches such as CS-SVM, DT, NB that yields the prediction [6]. To predict the risk factor of dengue fever from the health care records, the data set of dengue fever patients is taken from different hospitals. The structured data set comprises 820 records of patients and have 20 label attributes. The label attribute is the target class which has nominal values. This data has many values coming out from lab results, X-ray , the medical history and other related information. The results of the examination are taken for the feature selecting process that includes the pulse rate, the pressure on blood and other laboratory test results. The reports that constitutes on the data about the WBC count,, the count of platelet , Hematocrit test,

Table 1. Health care big data

DATA CATEGORY	ITEM	DESCRIPTION
STRUCTURED DATA	Patient Details	Patient age, Sex, Height, Weight etc.
	Habitual Details	Records of genetic details, smoking, drinking habits etc.
	Examination Reports	Blood checkup reports, Bp reports etc.
	Predicted Disease Reports	Disease records such as diabetics, Blood pressure etc.
	Pill Prescription Details	Kind of medicine followed by patient's details.
UNSTRUCTURED TEXT DATA	Patients readme illness	medical history
	Doctors Details	Doctor interrogation details
	Medical appointment details	How many times the patient's enrolled for medical checkup.

The Navie – Bayes classification is the one that is based on the probability.

This requires calculation of the probability in the attributes of the selected features. The formula for the condition probability for the estimation of features that are discrete along with the Gaussian ditribututive function for the estimation of attributes.

The patient data set is randomly spilt up into training data set and test data of the ratio of 5:1. The map reduce algorithm is implemented to run the classification in parallel fashion.

B.ML Algorithms

Improved naïve Bayesian classifier

It is a method of classification [7] based on Bayes' theorem with statement of unconventionality between prediction analysts. A Naive Bayes classifier accepts that particular feature in a class is dissimilar to any other feature. An NB classifier usually considers all the properties of the features in an independent manner and will contribute towards the probability in order to perform predictions.

The proposed improved naïve Bayesian classifier (INB) that classify the data sets by allowing the representation of dependencies among subsets of attributes.

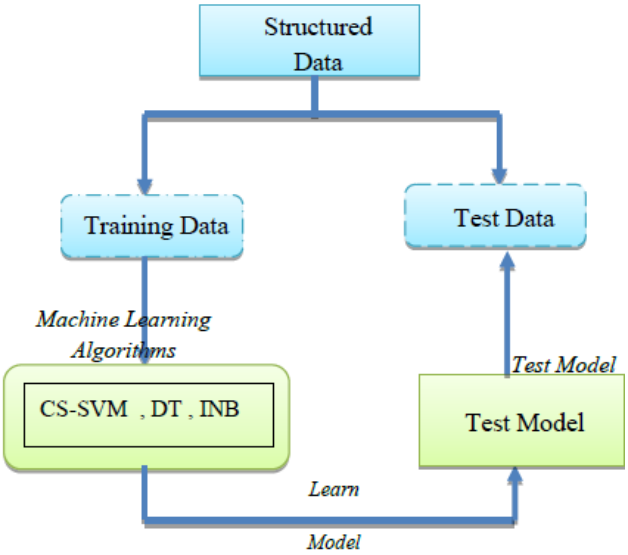

Fig:1 Prediction of structured data using Machine Learning algorithms.

The probability parameters P(c|x) from P(c), P(x) and P(x|c) factors Where P(c|x) is the posterior probability of predict class with the known predictor attribute. P(c) indicates the prior probability of class. P(x|c) is the likelihood which is the probability of predictor of given class and P(x) is the prior probability of predictor.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

The computation of posterior probability using NB is the outcome of prediction. Improved Naive Bayesian model provide high accuracy because the health care data attributes values are independent. It is a statistical model and provides high accuracy. This approach assumes all attributes are independent according to each other. This classifier can perform well in healthcare either by pre-processing or without pre-processing. The technique is very effective on a large range of complex problems.

C. Decision Tree

It is a supervised learning methodology that has three vital splitting conditions of data [5]. The information gain is the most important condition among the three. For a problem that is associated with multi class, the definition of entropy is shown in the equation below where D represents the training data.

$$E(D) = \sum_{x=1}^{n} P_x \log_2 P_x$$

The entropy is calculated when the number of information is zero, the information reaches the maximum. This measure is should be taken when multi-staged decision made.

D. CS-SVM

SVM is a statistical learning method, that maps the data in to high dimensional space F using non-linear mapping function $\varphi(X)$. The classification is progressed in high dimensional space by linear regression.

$f(x) = \omega \cdot \phi(X) + b = 0$

Where b is the threshold parameter and ω denotes the weight vector. Cuckoo Search algorithm is efficient method for optimization and it follows three basic principles. All the species of the cuckoo are involved in developing their own technique for the expansion of the chance of Hatching its eggs [6].

CS has three vital rules

1. Every cuckoo present in the nest will lay an egg in any moment that are predicted in a given time and then dumps the same in a nest on a random basis.
2. Only the quality eggs will proceed to the upcoming generation.
3. A few number of hosts that are more significant are then settled and the birds that lays the eggs are found using the host bird that has a probability

$P_a \in (0, 1)$.

This algorithm optimized to perform better solutions for employing the birds in the nested host. It is clear that CS searching algorithm often refreshes the nest of the birds with the best search path location, and it is represented as ,

$$X_i^{t+1} = X_i^t + \alpha \oplus \text{Levy}$$

The classification of CS-SVM is attained by collecting the training data sets and it is initialized to probabilistic parameter $P_a = 0.75$ and generated random nest N location. The training set is allocated to cross validation and then to find the current optimal location. The test set is obtained with SVM where the classification model is constructed to predict from the test set.

IV. ELM BASED DISEASE RISK PREDICTION CLASSIFICATION ALGORITHM

For handling high speed of data, Extreme Learning Method (ELM) [10] introduced to provide quicker learning speediness, great performance and with less computational complexity. The data present in health are kind of unstructured form when pointing to extreme machine learning even considering LM implementation including data platform. ELM learning is feed forward network and it modifies the sum of the input and the output in any intelligent method without the data pattern being retrained[9]. The challenge on Scalability is a vital aspect as any business is considered since it has the capacity to diverse the intention of the core logic at any moment.

A distributed system that is scalable is presented for the prediction of data which resulted in increased performance with minimum time complexity. Considering the unstructured data, the proposed ELM is designed as scalable learning system for supervised learning. The system that is trained will be able to predict the output which is certainly not known based on the input.The initial challenge when LM is implemented is the pre-processing of data which includes procedures such as normalization, missing value handling, data transformation, extraction of features and data training. The prime concept of data analytics is to extract useful information from data by identifying all the possible relations among the data.

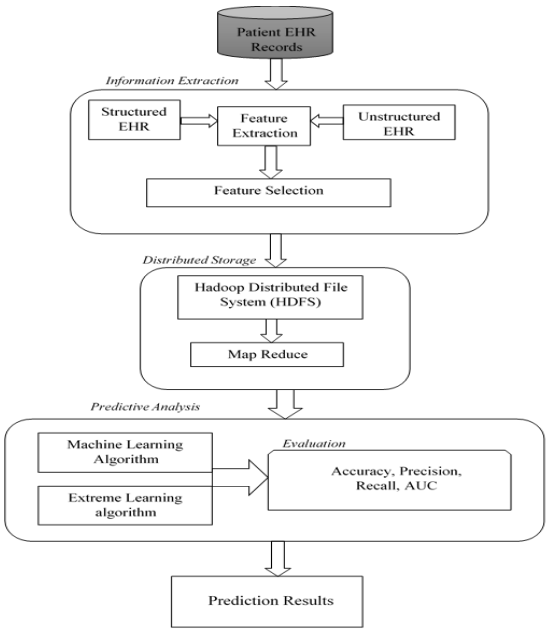

Fig 2. Overall Process of proposed block diagram

Other learning methods are affect the performance to decrease the level of gained performance measures foe each machines and also its affects the real time training and prediction. So as compared to these methods our proposed system is shows very good performance to data sets in a moderate size.

A.ELM feature Mapping

The output function of the ELM network structure for generalized Single hidden layer feed forward neural networks is calculated by the function,

$F(x) = \sum_{i=1}^{L} \beta i \, h_i(x) = h(x) \, \beta$

Where $\beta = [\beta_1, \ldots, \beta_l]^T$ indicates the output weight vector between the hidden layer and the output layer with m > 1 output nodes. $h(x) = [h_1(x), \ldots, h_L(x)]$ indicates the output vector of the hidden layer which can be termed as nonlinear ELM feature mapping classifier. Risk management and calculating potential risk cause.

V. RESULT AND DISCUSSIONS

The proposed methodology of standard optimiation in ELM can be extended in a linear fashion towards CS-SVM that has fewer optimization constraints and also it reduces the implementation overhead. Each dataset was applied to different classifier algorithms, they are Neural Network, Logistic Regression, Random Forest, Naive Bayes, and Cuckoo Search-Support Vector Machine. The statistical measures for various classification algorithms are discussed in Table 2.

Table 2. Performance measures for various ml algorithms

Classification Approach	Precision	Recall	Accuracy	AUC
Neural Network (NN)	85.2	88.1	83.8	91.9
Random Forest	67.3	94.3	67.3	88.1
SVM	85.8	86.7	82.4	92.4
Decision Tree(DT)	87.9	81.4	87.1	88.3
Logistic Regression	83.9	91.2	81.2	86.9
Naïve Bayes	91.2	95.4	88.6	95.5
Improved Naïve Bayesian	92.3	97.3	92.4	98.3
CS-SVM	93.6	92.5	93.5	97.9

The potential measures applied for measuring the predictive analytics are True Positive (TP), False Positive (FP), Positive Prediction (PP), and accuracy factor. The proposed work reaches the accuracy of 99%. The extended learning machine and machine learning classifiers are combined as a test data set and it is applied on 2 GB, 6 GB, 12 GB data sets are placed in a high performance distributed network. This requires high performance computing platforms and algorithms.

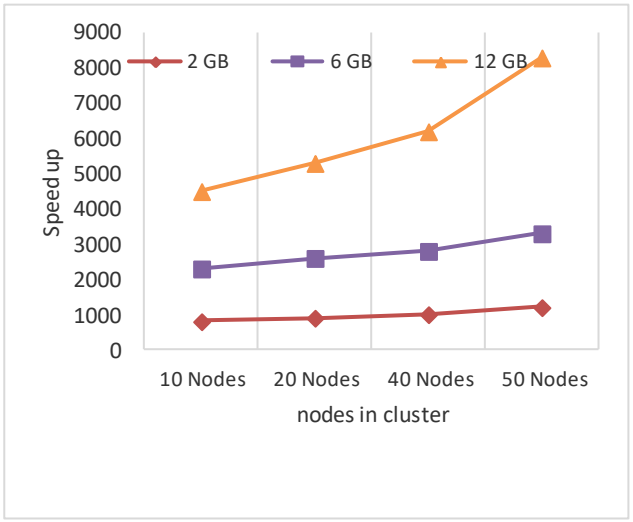

Fig.3 Speedups for each data sets

Figure 3 depicts the speed up in distributed framework that each cluster contains the considerable amount of data sets where the system attains greater speedups. The size of the hospital data sets increases, map reduce provides more success rate because of their higher parallelization.

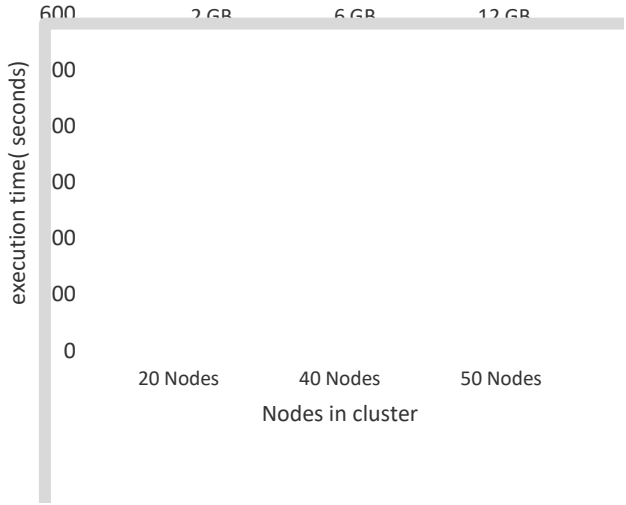

Fig:4 Execution Time for each data sets

Figure 4 depicts the time taken for execution in each test data sets that calculates the execution time based on the cluster series that are executed in various numbers of nodes. The execution time extends almost linearly with the input data set size. The proposed method helps to reduce the execution time to achieve the efficiency of this system.

VI. CONCLUSION

ELM is current and future trend in extreme large data sets. This paper focuses the accurate prediction of diseases using both structured and unstructured data sets. The data storage and the solutions would rather provide an better solution in the contrast of the storage mechanisms that are used traditionally. The research can be extended for a better results by enhancing the algorithm and services. Also proposed work using map reduce for unbalanced data are focused to attain scalability, fault tolerance in distributed platform. By the combination of these two type of data, the value of accuracy reaches 98.70%, so as to better evaluate the risk of dengue fever.

REFERENCES

1. Muhammad Umer Sarwar, Muhammad Kashif Hanif, Awais Mobeen, and lkjx]=09876543
2. I.Khan";lkjn bvAnalytics in Healthcare" International Journal of Advanced Computer Science and Applications, Vol. 8, No.6, PP 355-359, 2017
3. J. Han and M.Kamber , "Data Mining Concepts and Techniques" 2nd Edition, USA: Morgan Kaufinann pub.,2011.
4. Ma, L. Gu, B. Li, Y. Ma and J. Wang, "An Improved K-means Algorithm based on MapReduce and Grid", International Journal of Grid Distribution Computing, vol. 8, no. I, pp. 189-200,2015.
5. Dini Rahmawati and Yo- Ping huang, "Using C - support vector classification to forecast dengue fever epidemics in Taiwan" International Conference on System Science and Engineering, pp. 1-4, 2016
6. Zhengo Yao, Peng Liu, Lei Lei and Junjie Yin, "R- C 4.5 Decision Tree Model and its Applications To Health Care Data set" International Conference on Services Systems and Services Management, Vol. 2, PP. 1099 – 1103, 2005
7. Boris Milovic and Milan Milovic, "Prediction and Decision Making in Health Care using Data Mining" International Journal of Public Health Science (IJPHS) Vol. 1, No. 2, pp. 69 – 78, Dec 2012.
8. S.B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica 31, pp 249-268,July 2007.
9. Veronica S. Moertini and Liptia Venica , "Enhancing Parallel K-means using Map Reduce for discovering knowledge from big data", IEEE International conference on Cloud Computing and Big data Analysis, pp.81 – 87, 2016
10. J. Xin, Z. Wang, L. Qu, and G. Wang, "Elastic extreme learning machine for big data classification", vol.149, pp. 464 – 471, 2015.
11. Zhiqiong Wang, Junchang Xin, Hongxu Yang, Shuo Tian, Ge Yu, Chenren Xu, and Yudong Yao, " Distributed and Weighted Extreme Learning Machinenfor Imbalanced Big Data Learning" Tsinhua Science and Technology, vol. 22, pp. 160 – 173, 2017.