Implementing Machine Learning Techniques with Hadoop in Healthcare

Priti Sadaria[1], Achyut C. Patel[2]
Department of CS & IT - Atmiya University, Department of Statistics - Saurashtra University
[1]priti.sadaria@atmiyauni.ac.in
[2]acp2809@gmail.com

***Abstract*— In the era of technology, data is generated at high speed and processing and analyzing these very big data is a not easy task. Traditional database management systems are time consuming and are not able to analyze data fully. Healthcare industry have large amount of data but it is lack of analysis so hidden pattern cannot be identify for prediction of any diseases. To overcome such issue, Big Data is used to handle and control large volume of data which may be either in structured or in unstructured form. The hidden pattern can be identified and prediction can be made about future condition. Hadoop MapReduce has the capability to facilitate healthcare industry to get better prediction of diseases and make faster and proper judgment for right future treatment of patient by analysing healthcare data. Machine Learning algorithms can help to design predictive healthcare model for community wellness. I proposed system architecture to process and analyze data using Hadoop MapReduce with Machine Learning Techniques and ultimately it leads to prediction of diseases. As a result it increases life span as well as it leads to healthy life and reduce the rate of death by providing timely treatment.**

***Keywords*— Hadoop, Big Data, MapReduce, Healthcare, Machine Learning Technique**

I. INTRODUCTION

The research paper focus on developing Hybrid - SVM algorithm and model for Big Dataset processing by using the classification and clustering algorithms together. In my research I have focused on diabetes related disease because in this era this disease is spreading more quickly due to inappropriate life style and food habits. Ultimately it leads to chronic diseases like Nephropathy, Retinopathy and Cardio Vascular Diseases.Because of this strong requirement arise to develop a new algorithm for the prediction purpose of diseases. The research focus on developing Hybrid - SVM algorithms for predicting diseases.

II. TOOLS AND TECHNIQUES

In the research following tools and techniques are used.

A. Hadoop MapReduce
B. K - means clustering with MapReduce
C. Support Vector Machine

A. Hadoop MapReduce

Hadoop works with the concepts of HDFS and MapReduce. HDFS is used to store a huge dataset in form of file and the dataset can be process by using MapReduce technique.

B. K - means clustering with Hadoop MapReduce

K -means clustering is very useful algorithm to identify clustering pattern among the dataset in healthcare sector [1]. The Figure 5.1indicates coding of K - means clustering and as a result there are two categories can be identified, diabetic and non- diabetic. Diabetic cluster further distributed into three categories Low, Medium and High.

C. Support Vector Machine (SVM)

Pattern identification is an essential task in Machine Learning and it is known as classification technique. Prediction function f can be constructed by using supervised learning algorithm with implementing training data [2].

Here classification can be achieved by using N-Dimensional hyper plane and each point is indicated as a data item. Hyper plane is a line which linearly separates and classifier a set of data. Co-ordinates of each data items can be represent in the form of support vectors which are nearest to the hyper plane. Here margin represent the distance between the hyper plane and the closest data point from one of the set. Finally optimal

hyper plane distinguish two categories of clusters depending on target variable, one is on one side of hyper plane and the other is on other side.

III. PREPARING DATA SET

In the research I have used PIMA Indian diabetic data set which includes following attributes:

A. Attributes utilized for clustering

- Life style – Habit of eating and physical activity
- Diet – Specify food habit
- Exercise
- Stress
- Family History
- Obesity
- Hypertension
- Glycosylated haemoglobin(HbA1C)
- Area into which patient belong

K - means clustering technique can be used along with these all attributes to categorize data into two category, diabetic cluster and non-diabetic cluster in first level and diabetic cluster again categorize into three cluster, Low Diabetic, Medium Diabetic and High Diabetic.

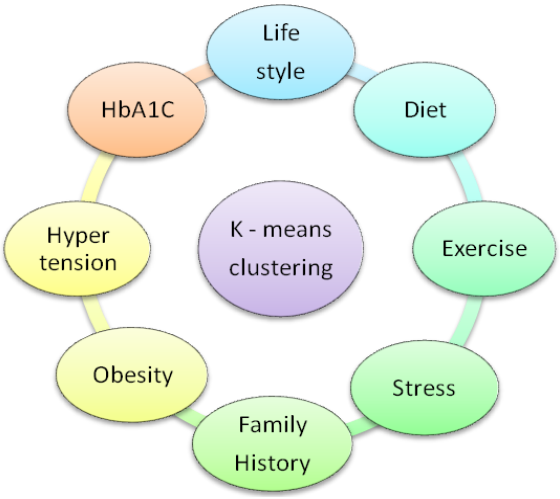

Fig. 1 Attributes Utilized for Clustering

B. Clinical attributes utilized for classification for prediction with SVM

- Number of years with diabetes
- Creatinine
- High Density Lipoprotein - HDL
- Low Density Lipoprotein - LDL

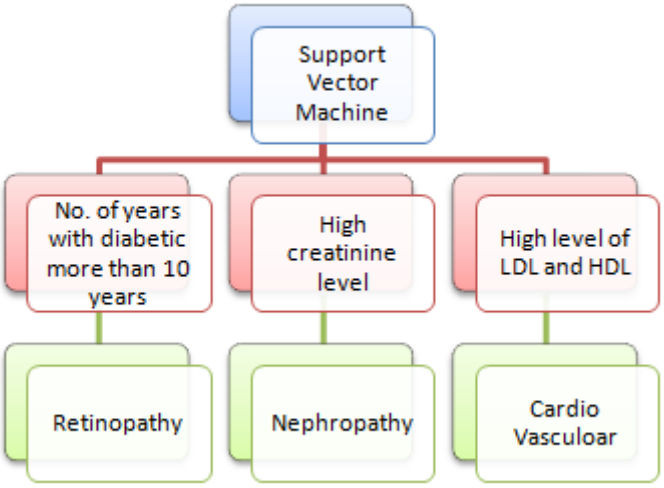

Fig. 2 Clinical Attributes Utilized for Classification

Classification by SVM carried out depending on the clinical attributes No. of years more than 10 with diabetic, Creatinine, LDL and HDL, as a result prediction also can be done for diabetic complicated diseases.

IV. PERFORMANCE OF HYBRID - SVM ALGORITHM WITH MAPREDUCE

In the research, Hadoop MapReduce used with K - means clustering technique and SVM for prediction. The input data transferred from local machine to HDFS, The algorithm distributes the data into n number of cluster by using K - means clustering technique and the mapper function starts to process the data. Generated clusters are provided as input to Hybrid SVM to build the prediction model.

Hybrid - SVM algorithm steps:

1. Inputted data read from HDFS
2. Mapper receives input data
3. Now initialized the center point for the cluster and point the nearest cluster for each data point.
4. Next set the position of each cluster to the mean value of all data points.
5. Iterate step 3 and 4 still to get union.
6. Once union achieved, classifiers are build to create a model by using training data set over cluster outputs.
7. Prediction can be done on new model by using test dataset.
8. At last process end and prediction output file is generated.

The research is helpful for prediction of diabetic patients who may have risk of occurring serious diseases like Cardio Vascular Diseases, Nephropathy and Retinopathy.

Here the dataset is processed into two steps, one is mapper and the other is reduce. As Figure5.5 indicates, first dataset is loaded on HDFS and latter on data is provided to mapper. K -means clustering is used by the mapper and as a result numbers of clusters are created. Hybrid - SVM uses clustered data provided by the mapper in key-value pairs and shuffling process is performed. At last, reducer aggregated all values and saved the result on HDFS in the form for the file.

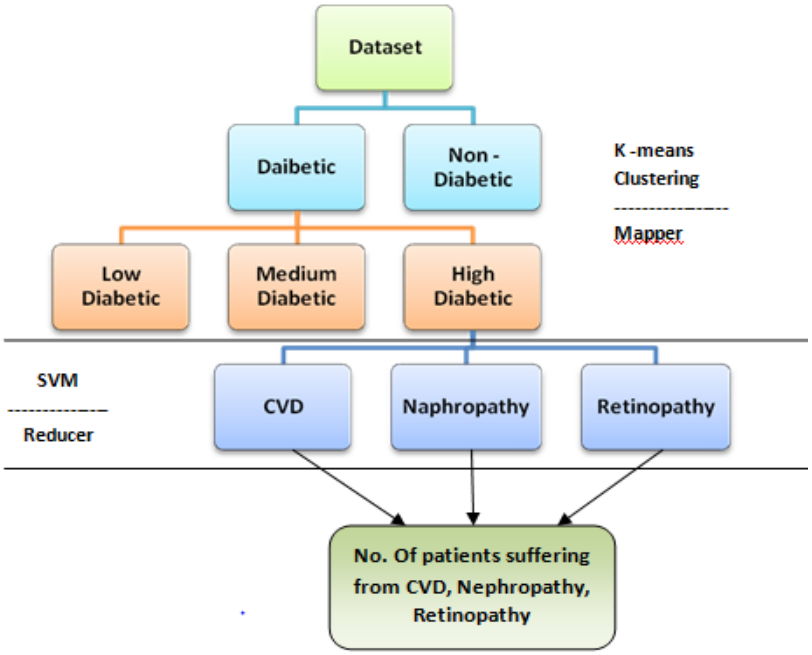

Fig. 3 Working Scenario in Hybrid - SVM for Clustering and Classification

Six datasets having size 650000, 843000, 857229, 1650000 and 6000000 were analyzed by implementing Hybrid-SVM methodology and processing time was evaluated. From the research it is concluded that for large dataset, stand alone mode was not competent process the task effectively because of resource constraints. The task effectively completed using Hybrid-SVM algorithm for the district Baroda is shown in Fig. 3 indicates that the dataset categorized into two categories depending on clustering, diabetic and non diabetic. Diabetic cluster again distributed into three categories low risk, medium risk and high risk.

Table I
Execution time for processing of six datasets Using Hybrid-SVM in standalone mode with Hadoop

Dataset	Districts	No of Records	Processing Time
1	Anand	650000	102
2	Rajkot	843000	109
3	Baroda	857229	120
4	Ahmedabad	1650000	187
5	Surat	2000000	228
6	Total of 5 District	6000000	Not able to process

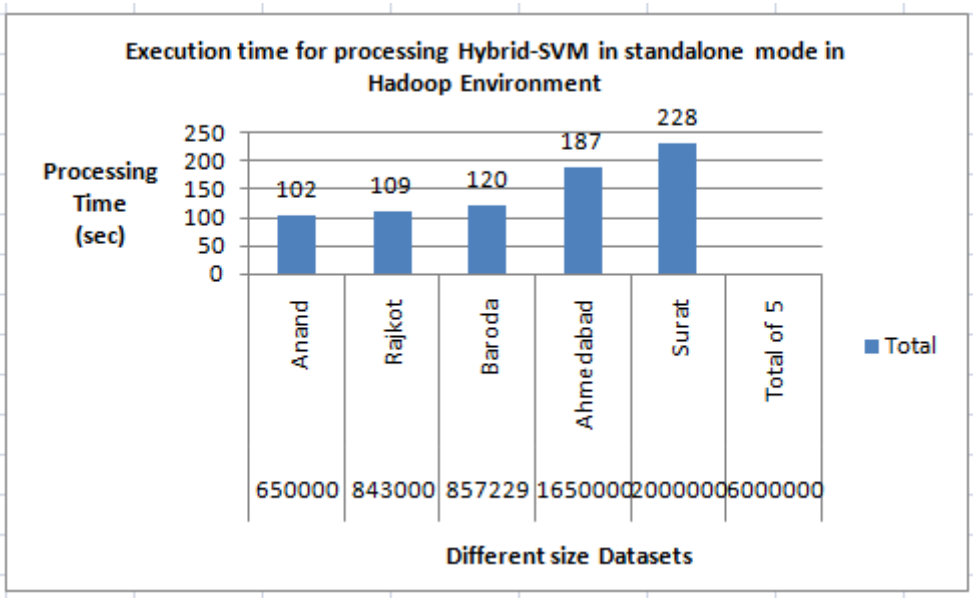

Fig. 4 Execution time for processing Hybrid-SVM in standalone mode in Hadoop Environment

Table I and Fig. 4 indicate that the processing time was relatively raised according to the size of different datasets when we analyzed using Hybrid-SVM algorithm. Table I shows time taken for 650000, 843000, 857000, 1650000 and 2000000 records was 102, 109, 120, 187 1nd 228 seconds respectively. The system was not competent to process the dataset when we provide total five districts 6000000 records as input to the system. It happens due to resource constraints in standalone mode.

V. HYBRID-SVM TECHNIQUE FOR THE PREDICTION OF DIABETIC DISEASES

The dataset file was loaded into HDFS and then mapper and reducer functions work on that. The inputted files are split into number of blocks by using new techniques, K-means and SVM. At initial level classification, the dataset is clustered by considering the risk factors like obesity, food habits, family history, exercise, HBA1C etc. Score value is assigned to each risk factor, for example 5 points are assigned to family history and each remaining factors assigned to 1 point. Depending on the risk factor value, for individual records total score was saved into output variable. The output variable categorize into two clusters, diabetic and non-diabetic. The value of output variable was among 1 to 5 and for non-diabetic cluster and among 6 to 10 for diabetic cluster.

The diabetic cluster again clustered into three groups depending on the score value. If the score value is 9-10 then it is categorized as diabetic high risk, if the score value is 8 then it categorized as diabetic medium risk and if the score value is 6-7 then it categorised as diabetic low risk. In the second level, for prediction of diabetic complication again the diabetic high risk cluster is divided into three groups: CVD, Nephropathy, and Retinopathy depending on LDL, HDL and Creatinine level and more than 10 years.

Here SVM supervised learning technique is used for diabetic cluster and by processing it intermediate key is generated. By using the reduction function data in each section is processed and intermediate values are combined and the generated result saved into the output files in the Hadoop HDFS.

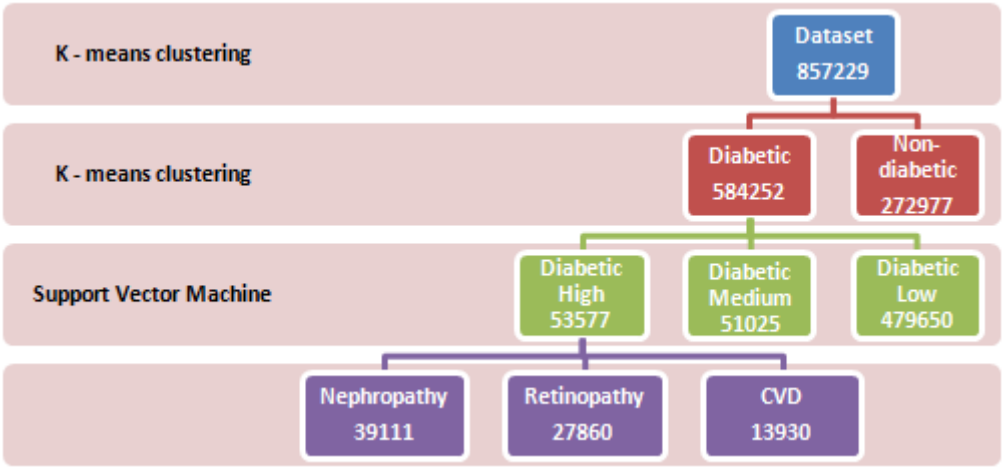

Fig. 5 Hybrid – SVM Technique to Predict the Diabetic Related Diseases

The Fig. 5 indicates the clustering of dataset with records 857229. At first level dataset has been clustered into Diabetic and Non-Diabetic category. Diabetic category includes 584252 records and Non-Diabetic category includes 272977 records. In next step, the Diabetic cluster again distributed into three cluster: Diabetic High risk having 53577 records, Diabetic Medium risk having 51025 records and Diabetic Low risk having 479650 records. By applying SVM machine learning technique on Diabetic High risk cluster, prediction of diabetic related diseases Nephropathy, Retinopathy and CVD can be done.

The Table II shows percentage ratio for individual category of diabetic related diseases.

Table II
Diabetic Related Diseases for Diabetic High Risk Patients

Disease Name	Total patients	Percentage Ratio
Nephropathy	39111	73%
Retinopathy	27860	52%
CVD	13930	26%

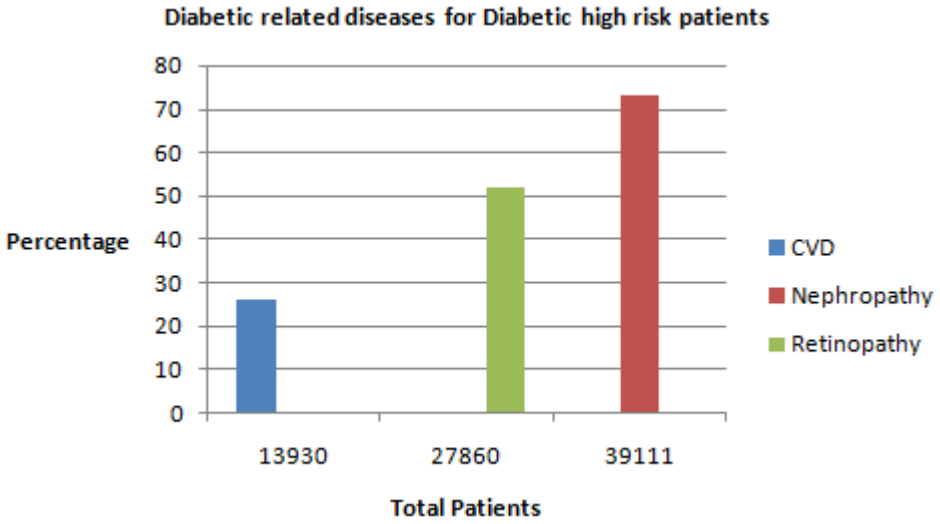

Fig. 6 Diabetic Related Diseases for Diabetic High Risk Patients

VI. CONCLUSIONS

In my research, the Hybrid-SVM algorithm has been used for prediction of diabetic related diseases. By using Hybrid-SVM algorithm in Hadoop environment, we can predict diabetic related diseases like CVD, Nephropathy and Retinopathy by processing large dataset at high speed. When the dataset size is larger than 6 million, the processing of dataset is not possible using Hybrid-SVM in Hadoop standalone

mode so strong requirement arise to use Cloud Computing Environment for processing data in parallel in distributed form.

The research shows that processing time can be decreased by using Hybrid-SVM algorithm which is developed by combination of MapReduce, K-means and Support Vector Machine. Here clustering has been done using K-means clustering algorithm and the classification and prediction has been performed using SVM.

References

[1] Wullianallur Raghupathi, Viju Raghupathi, "Big data analytics in healthcare: promise and potential", Health Information Science and Systems, 2(3): 2-10. 2014.

[2] Padmavathi Jabardhanan, L.Heena, Fathima Sabika –" Effectiveness of Support Vector Machines in Medical Data mining", Journal of Communications Software and Systems 11(1):25-30 · April 2015

[3] Muni kumar N, Manjula R, " Role of Big Data Analytics in, Rural Helath Care – A Step Towards Svasth Bharath", International Journal of Computer Science and Information Technologies

[4] Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. IEEE J Biomed Health Inform. 2015

[5] Zhanquan Sun ―Study on Parallel SVM Based on MapReduce in conference on world comp. 2012.

[6] Ashwin Belle, Raghuram Thiagarajan, S. M. Reza Soroushmehr Fatemeh Navidi, Daniel A. Beard, and Kayvan Najarian, Big Data Analytics in Healthcare BioMed Research International Volume 2015, Article ID 370194

[7] Dr. Saravana kumar N M , Eswari T , Sampath P & Lavanya S, "Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing(ISBCC'15)