International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

Machine Learning Based Method for Prediction of Heart Disease in Big Data Environment

Sharmila Rengasamy, Chellammal Surianarayanan, Pethuru Raj Chellaih

Abstract: ***Prediction of diseases is one of the challenging tasks in healthcare domain. Conventionally the heart diseases were diagnosed by experienced medical professional and cardiologist with the help of medical and clinical tests. With conventional method even experienced medical professional struggled to predict the disease with sufficient accuracy. In addition, manually analysing and extracting useful knowledge from the archived disease data becomes time consuming as well as infeasible. The advent of machine learning techniques enables the prediction of various diseases in healthcare domain. Machine learning algorithms are trained to learn from the existing historical data and prediction models are being created to predict the unknown raw data. For the past two decades, machine learning techniques are extensively employed for disease prediction. Despite the capability of machine algorithm on learning from huge historical data which is stored in data mart and data warehouses using traditional database technologies such as Oracle OnLine Analytical Processing (OLAP). The conventional database technologies suffer from the limitation that they cannot handle huge data or unstructured data or data that comes with speed. In this context, big data tools and technologies plays a major role in storing and facilitating the processing of huge data. In this paper, an approach is proposed for prediction of heart diseases using Support Vector Algorithm in Spark environment. Support Vector Machine algorithm is basically a binary classifier which classifies both linear and non-linear input data. It transforms the non-linear data into hyper plan with the help of different kernel functions. Spark is a distributed big data processing platform which has a unique feature of keeping and processing a huge data in memory. The proposed approach is tested with a benchmark dataset from UCI repository and results are discussed.***

Keywords-Support Vector Machine for Heart disease machine; Spark MLLib for heart disease prediction; big data for disease prediction; machine learning algorithms for disease predication.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
Sharmila Rengasamy*, Department of Computer Science, Bharathidasan University Constituent Arts & Science College, Navalurkuttapattu, Tiruchirappalli-620027 TamilNadu, India, Email: sharmiparam@gmail.com
Chellammal Surianarayanan, Department of Computer Science, Bharathidasan University Constituent Arts & Science College, Navalurkuttapattu, Tiruchirappalli-620027 TamilNadu, India, Email: chelsganesh@gmail.com
Pethuru Raj Chellaih, Site Reliability Engineering (SRE) Division, Reliance Jio Infocomm. Ltd. (RJIL), AVANA Building, Iblur Village, Sarjapur Road, Bangalore 560103, peterindia@gmail.com..

I. INTRODUCTION

Data mining is a field of computer science which extract hidden and useful knowledge from huge amount of data [1-2]. Data mining combines the techniques from various disciplines such as statistics, machine learning, artificial intelligence, computing analytics, text analytics, database technologies, etc. In healthcare domain, it is a routine activity that various data such as patient data, clinical data, medical data such as scan images, Electro Cardio Gram(ECG), Electro Encephalo Graph(EEG), Ultra Sound scan, Computed Tomography(CT) scans, Magnetic Resonance Imaging(MRI) scans, prescription notes, etc. The above data is being archived over decades. Certainly drawing inference from the archived data is tedious and time consuming. But it is definitely needed to explore the archived data to find out the hidden and interesting knowledge which helps in better decision making. In this context, data mining techniques play a critical role analyzing the archived data by using different algorithms such as clustering, classification, prediction, pattern recognition, regression, etc

Data mining is really a boon to health care domain as it provides various descriptive, predictive and prescriptive mining tasks which when applied on the historical data, yields many meaningful knowledge such as characterization of data, summarization of data, finding associations between different data elements, etc. More specifically, machine learning algorithms[3] such as Naive Bayes, Decision Tree, Support Vector Machine, Artificial Neural Networks etc are training to learn from huge training data set and to create a set of classification rules called classifier or model. The model is validated for its accuracy with a collection of test data. Once the accuracy is found to be sufficient, the model will be put into practice to predict or classify the new or unknown data

For performing machine learning, the historical data is required to be stored in data marts and data warehouses using conventional database technologies such as Oracle OnLine Analysis Processing (OLAP). But the conventional database technologies suffer from the limitation that they become inadequate when the data size is of the order of Terra Bytes(TB). In such situations big data techniques and tools play a critical role by providing storage and processing platforms where the massive data of size exceeding TB can be easily stored in big data solutions such as Hadoop, Spark, etc[4-9]. In addition, latest tools such as Apache Spark is efficient in keeping huge amount of data in memory itself and handle it in almost real time by using micro batch processing techniques Thus in this paper, it is proposed to employ Apache Spark,

a distributed big data processing platform which facilitates the in-memory processing of historical data in order to enhance the performance of machine learning algorithm.

It is proposed to use Support Machine from Spark Machine Learning Library for prediction of heart diseases. The proposed approach is evaluated with a test set collected from UCI repository. The results are evaluated using different evaluation measures such as precision, recall and f-measure

The paper is organized as follows. Section II highlights various research works that are related to the theme of the proposed work. Section III describes the proposed approach. Section IV presents the evaluation of proposed method with an overview of dataset used for experimentation. Section V concludes the paper with directions for future research.

II. RELATED WORK

There are significant amounts of research works that use machine learning algorithms for prediction of heart diseases. This section gives an overview about how different machine learning algorithms are being used for prediction of heart diseases. In [10], the authors reviewed different machine learning techniques such as Support Vector Machine(SVM), Decision Tree, K-Nearest Neighbour, etc., that are used for heart disease prediction. The authors found that SVM classifies the data set by finding best hyper plane that maximizing the margin between two classes. In[11], authors reviewed different machine learning classification techniques such as Naive Bayes(NB), Support Vector Machine(SVM), Artificial Neural Network(ANN), Radial Basis Function(RBF), Decision Tree(DT), K-Nearest Neighbour(K-NN) and Genetic algorithm for heart disease prediction.

The review discusses that the accuracy of a classifier gets influenced by various factors such as tool used analytics, dataset, number of attributes, number of records in the dataset and classifier used for prediction. In[12],authors have reviewed various research works which have been used for heart disease prediction. In[13], the authors performed some survey on how different machines learning algorithms are used for heart disease prediction. In[14], an approach is proposed for heart disease prediction using Machine Learning Algorithms which are optimized by Particle Swarm Optimization(PSO) and Ant Colony Optimization(ACO) techniques. In [15], the authors performed experimentation with five different classifiers, namely, namely Artificial Neural Network, Support Vector Machine, Random Forest, Decision Tree, K-Nearest Neighbour and found that Random Forest produces highest accuracy. In[16], authors used evolutionary rule learning for heart disease prediction. In[17],authors designed HPPS(Heart Problem Prediction System) using machine learning keeping in mind nine different risk factors, namely, family history, smoking, hypertension, dyslipedimia which refers to a high level of lipids like cholesterol, triglycerides, fasting glucose, life style, Coronary Artery Bypass Grafting(CABG) surgery and high serum. In this research work, the SVM-RBF gives better performance. In [18], an application has been developed to detect the vulnerability of a heart disease using given symptoms like age, sex, pulse rate etc., by using Multi Layer Perceptron(MLP). In[19], a novel method , Hybrid Random Forest with a Linear Model (HRFLM) has been proposed to find important features in order to improve accuracy in heart disease prediction. In[20], how different big data techniques can be used for prediction of heart diseases have been discussed. In contrast to the above literature, in this work, it is proposed to analyze the prediction of machine learning algorithms in big data environment.

III. PROPOSED WORK

The block diagram of the proposed approach is shown in Figure 1.

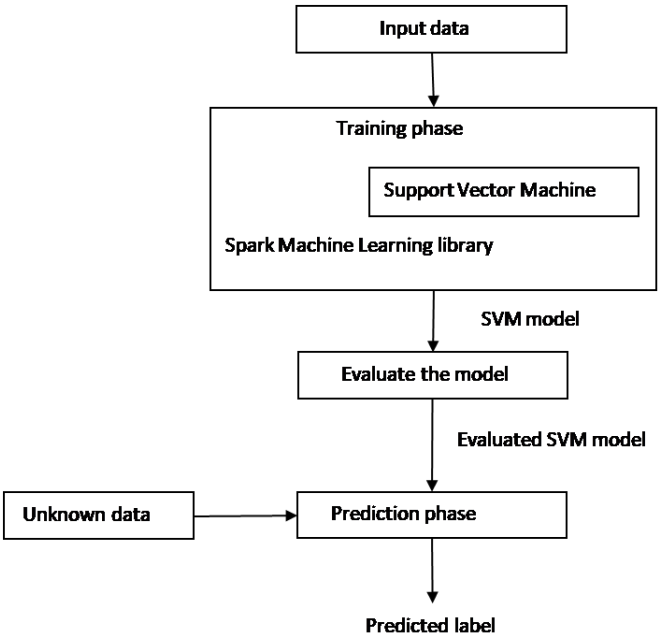

Figure 1. Block diagram of the proposed approach

In the proposed approach input data is collected from UCI repository and the data is pre processed for its completeness and correctness.

Spark is a distributed big data processing platform which can keep huge data in memory. Spark offers a machine learning library which consists of different machine learning algorithms such as Support Vector Machine, Random Forest, Regression, clustering, etc. Spark supports different languages such as Scala, Python, Java, etc. In this work, the SVM algorithm in Spark MLLib is trained with the preprocessed input data. (SVM) is basically a binary classifier which builds hyper plane with margin of separation between both the class labels is maximized.

The algorithm transforms the original feature space to higher dimensional space with the help of kernel functions. The algorithm learns from different records of input data and constructs the model or classifier which represents a set of classification rules.

Then the constructed SVM model is evaluated for its performance using accuracy measure. If the accuracy of the model is found sufficient, then the model is put into prediction phase where it predicts the label of any unknown data given as input

IV. EXPERIMENTATION

A test collection of 303 records have been collected as input data from Cleveland data source of UCI repository. The collection contains 303 records since, 6 records which contain missing values are not considered for experiment. Each record contain 13 attributes, namely, age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal and one class label column(num). The data type, range of various attributes, are given in Table I

TABLE I. DESCRIPTION OF DATASET

S .No	Attribute Name	Attribute Type	Minimum	Maximum
1	Age	Numeric	29	77
2	Sex	Numeric	0	1
3	Cp	Numeric	1	4
4	Trestbps	Numeric	94	200
5	Chol	Numeric	126	564
6	Fbs	Numeric	0	1
7	Restecg	Numeric	0	2
8	Thalach	Numeric	71	202
9	Exang	Numeric	0	1
10	Olepeak	Numeric	0.0	6.2
11	Slope	Numeric	1	3
12	Ca	Categorical	0	3
13	Thal	Categorical	3	7
14	Num	Numeric	0	4

In the data set, as far as target or class label is concerned, there are 5 possible values. They are 0,1,2,3 and 4. Of the five values, 0 refers to a case with no heart disease. The remaining 4 values denote the presence of heart disease, In the proposed approach, the target class values having values 1,2,3 and 4 are taken simply as 1 so that there would be only two values, 0 and 1. Also, in the data set there are 137 records having class label 1(indicates the presence of heart disease) and 160 records having class label 0(indicates the absence of heart disease). A programme has been coded in Python using SVM library to create SVM model. Experiments have been carried out with to split ratios, 70:30 and 80:20. Since, the nature of dataset is not known the experiment is carried out with different kernel functions, *linear, polynomial, sigmoid* and Radial Basis Function (RBF). It is proposed to compute the confusion matrix for various kernel functions and compare the performance of the classifier for various kernel functions using 4 different evaluation measures, accuracy, precision, recall and f-score which are computed using the following formulae given through equations (1), (2), (3) and (4)

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$recall = \frac{TP}{TP + FN} \quad (2)$$

$$precision = \frac{TP}{TP + FP} \quad (3)$$

$$f - score = \frac{2(recall \times precision)}{(recall + precision)} \quad (4)$$

In the above equations, TP stands for *True Positive* which denotes an outcome corrected that is correctly predicted as positive class by the classifier, TN stands for *True Negative* which denotes an outcome that is correctly predicted as negative by the classifier, FP stands for *False Positive* which denotes a outcome incorrectly predicted as negative class by the classifier and FN stands for *False Negative* which denotes an outcome that is incorrectly predicted as negative class by the classifier.

V. RESULTS

Confusion matrix obtained using various kernel functions with two different split ratios(70:30 & 80:20) respectively is given in Table II.

TABLE II. CONFUSION MATRIX

Kernel	Actual/ Predicted class label	Split ratio = 70:30		Split ratio = 80:20	
		Predicted NO	*Predicted YES*	*Predicted NO*	*Predicted YES*
Linear	*Actual NO*	44	4	28	2
	Actual YES	12	30	8	22
RBF	*Actual NO*	44	4	28	2
	Actual YES	10	32	8	22
Polynomial	*Actual NO*	43	5	28	2
	Actual YES	14	28	10	20
Sigmoid	*Actual NO*	43	5	28	2
	Actual YES	11	31	8	22

As mentioned earlier, the test data consists for 297 records. With 70:30 split ratio, 30% of records i.e. 90 records would form as test data. Out of 90 records, 42 records have their class label as positive and remaining 48records have their class label as negative. Similarly, with 80:20 ratio, 20% of records, i.e. 60 records would form as test data. Out of 60 records, 30 records have their class label as positive and remaining 30 records have their class label as negative. From the confusion matrix given in Table II, TP, TN, FP and FN for different kernel functions are computed. From the values of TP, TN, FP and FN, the values of accuracy, precision recall and F-Score are computed and the computed values are given as classification report obtained using different kernel functions are given in Table III. From Table III, it is found that RBF kernel function gives the highest accuracy of 84% with split ratio, 70:30. In Table III, the precision, recall and F-Score values are shown for both positive and negative classes. Since, positive class is more significant for analysis, the values of precision, recall and F-Score for positive class are considered for comparison Thus, RBF kernel yields highest values of accuracy, precision, recall and F-Score as 84%, 0.89, 0.76 and 0.82 respectively.

TABLE III. CLASSIFICATION REPORT

Kernel Name	Split Ratio	Accuracy in %	Precision		Recall		F-Score	
			0	*1*	*0*	*1*	*0*	*1*
Linear	70:30	82	0.79	0.88	0.92	0.71	0.85	0.79
	80:20	83	0.78	0.92	0.93	0.73	0.85	0.81
RBF	70:30	84	0.81	0.89	0.92	0.76	0.86	0..82
	80:20	83	0.78	0.92	0.93	0.73	0.85	0.81
polynomial	70:30	78	0.75	0.85	0.90	0.67	0.82	0.75
	80:20	80	0.74	0.91	0.93	0.67	0.82	0.77
Sigmoid	70:30	82	0.80	0.86	0.90	0.74	0.84	0.79
	80:20	83	0.78	0.92	0.93	0.73	0.85	0.81

VI. CONCLUSION

An approach for prediction of heart diseases using SVM from Spark MLLib has been proposed in order to apply of in-memory concept of Spark. When data grows, Spark is capable of keeping huge data in memory and process the data efficiently. The proposed approach has been tested with 297 records. But to study the effect of in-memory concept of Spark, as a future it is proposed to create a huge data collection of around one million records. In addition, it is planned to split the data into different nodes and analyse impact of big data tools in the accuracy of prediction.

REFERENCES

1. Jaymin Patel, Prof.TejaUpadhyay and Dr. Samir Patel, "Heart Disease Prediction using Machine Learning and Data Mining Technique", IJCSC, Volume 7, pp 129-137, March 2016.
2. Vijayashree J and N.Ch. Sriman NarayanaIyengar, "Heart Disease Prediction System using Data Mining and Hybrid Intelligent Techniques: A Review", International Journal of Bio-Science and Bio-Technology, Volume 8, o.4, pp 139- 148, 2016
3. S. Bagavathy, V.Gomathy, S. Sheeba Rani, Sujatha .K, Bhuvana M.K, and Monica Murugesan, "Early Heart Disease Detection using Data Mining Techniques with Hadoop Map Reduce", International Journal of Pure and Applied Mathematics, Volume 119, No 12, pp 1915-1920, 2018.
4. Abderrahmane Ed-daoudy and Khali Maalmi, " Real-time machine learning for early detection of heart disease using big data approach", IEEE, 2019.
5. R. Venkatesh, C. Balasubramanian and M. Kaliappan, " Development of Big Data Predictive Analytics Model fior Disease Prediction using Machine Learning Technique", Journal of Medical Systems, 2019
6. Vinitha s, Sweetlin s, Vinusha H and sajini s, "Disease Prediction using Machine Learning over Big Data", Computer Science & Engineering: An International Journal, Volume 8, No 1, pp 1-8, 2018.
7. Shraddha Subhash Shirsath and Pro. Shubhangi Patil, " Disease Prediction using Machine Learning over Big Data", International Journal of Innovative Research in Science, Engineering and Technology, Volume 7, Issue 6, pp 6752-6757, 2018.
8. T.Nagamani, S.Logeswari and B. Gomathy, "Heart Disease Prediction using Data Mining with Mapreduce Algorithm", International Journal of Innovative Technology and Exploring Engineering, Volume 8, Issure 3, 2019
9. Javier Andreu-Perez, Carmen C.Y. Poon, Robert D. Merrifield, StephenT.C.Wong and Guang-Zhong Yang, IEEE journal of Biomedical and Health Informatics, Volume 19, No 4, 2015.
10. Himanshu Sharma and M A Rizvi, "Prediction of Heart Disease using Machine Learning Algorithms: A Survey", International Journal of Recent and Innovation Trends in Computing and Communication, Volume 5, Issue 8, pp 99-104, 2017
11. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE
12. Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982.
13. SP Rajamhoana and C. Akalya Devi and K. Umamaheshewari and R.Kiruba, " Analysis of Neural Networks based Heart Disease Prediction System", IEEE, pp 233-339, 2018.
14. V. V Ramalingam, Ayantan Dandapath and M. Karthik Raja, "Heart Disease Prediction using Machine Learning techniques: a survey", International Journal of Engineering & Technology, 7 (28), pp 684 – 687, 2018
15. Youness Khourdifi and Mohamed Bahaj, "Heart Disease Prediction and Classification using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization", International Journal of Intelligent Engineering & Systems, Volume 12, No 1, pp 242-251, 2019
16. Virender Ranga and D. Rohila, "Parametric Analysis of Heart Attack Prediction using Machine Learning Techniques", International Journal of Grid and Distributed Computing, Volume 11, No 4, pp 37-48, 2018
17. Aakash Chauhan, Aditya Jain, Purushottam Sharma and Vikas Deep, "Heart Disease Prediction using Evolutionary Rule Learning", International Conferrence on "Computational Intelligence and Communication Technology, pp 1-4, 2018
18. Nimai Chand Das Adhikari, arpana Alka and Rajat Garg, "HPPS: Heart Problem Prediction System using Machine Learning", pp 23-37, 2017
19. Aditi Gavhane, Gouthami Kokkula, Isha Pandya and Prof. Kailas Devadkar, "Prediction of Heart Disease using Machine Learning", IEEE Conference, pp 1275-1278, 2018
20. Senthil Kumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava, "Effective Heart Disease Prediction using Hybrid Machine Learning Techniques", IEEEAccess, Volume 7, pp 81542- 81554, 2019
21. Prema Jain and Amandeep Kaur, "Big Data Analysis for Prediction of Coronary Artery Disease", IEEE, pp 188-193, 2018

AUTHORS PROFILE

Sharmila Regnasamy pursed Master of Computer Applications from Bharathidasan University of India in 2002. She is currently pursuing Ph.D. and working as Lecturer in Department of Computer Science, Bharathidasan University Constituent Arts & Science College, Trichy, India. Her main research work focuses on the impact of Big Data techniques for prediction of diseases using machine learning algorithms. She has 12 years of teaching experience and 3 years of research experience

Chellammal Surianarayanan is an Assistant Professor of Computer Science in Bharathidasan University Constituent Arts & Science College, Tiruchirappalli, TamilNadu, India. She earned a doctorate in Computer Science by developing computational optimization models for discovery and selection of semantic services. She published research papers in Springer Service-Oriented Computing and Applications, IEEE Transactions on Services Computing, International Journal of Computational Science, Inderscience& SCIT Journal of Symbiosis Centre for Information Technology, etc. She produced book chapters with IGI Global, CRC Press. She has been a life member in professional bodies such as Computer Society of India, IAENG, etc.. Before coming to Academic service, Chellammal Surianarayanan served as Scientific Officer in Indira Gandhi Centre for Atomic Research, Department of Atomic Energy, Government of India, Kalpakkam, TamilNadu, India. She was involved in the research and development of various need-based development of embedded systems and software applications.

Her remarkable contributions include the development of an embedded system for lead shield integrity assessment system, portable automatic air sampling equipment and the embedded system for detection of lymphatic filariasis. Totally she has 22 years of academic and industrial experience

Pethuru Raj Chellaih has been the chief architect and vice president of the Site Reliability Engineering (SRE) Center of Excellence (CoE) division, Reliance Jio Infocomm Ltd. (RJIL), Bangalore. His previous stints are in IBM Cloud center of Excellence (CoE), Wipro consulting services (WCS), and Robert Bosch Corporate Research (CR). In total, he has gained more than 17 years of IT industry experience and 8 years of research experience. Finished the CSIR-sponsored PhD degree at Anna University, Chennai and continued with the UGC-sponsored postdoctoral research in the Department of Computer Science and Automation, Indian Institute of Science, Bangalore. Thereafter, he was granted a couple of international research fellowships (JSPS and JST) to work as a research scientist for 3.5 years in two leading Japanese universities. Published more than 30 research papers in peer-reviewed journals such as IEEE, ACM, Springer-Verlag, Inderscience, etc. He has authored 10 books thus far and focus on some of the emerging technologies such as IoT, Cognitive Analytics, Blockchain, Digital Twin, Docker Containerization, Data Science, Microservices Architecture, fog/edge computing, etc. He has contributed 30 book chapters thus far for various technology books edited by highly acclaimed and accomplished Professors and professionals.