International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 3 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT195350

Chronic Diseases Prediction over Bigdata by using Machine Learning

Shreekanth Jogar[*1], Pavankumar Naik[*1], Veeramma Vyapari[2], Madevi Vaddar[2], Kavita Dambal[2], Bheemavva Hatti[2]
[*1]Assistant Professor Department of Computer Science and Engineering, SKSVMACET, Laxmeshwar, Karnataka, India
[2]Student Department of Computer Science and Engineering, SKSVMACET, Laxmeshwar, Karnataka, India

ABSTRACT

With big data growth in biomedical and healthcare communities, accurate analysis of medical data benefits early disease detection, patient care and community services. However, the analysis accuracy is reduced when the quality of medical data is incomplete. Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks. In this paper, we streamline machine-learning algorithms for effective prediction of chronic disease outbreak in disease-frequent communities. We experiment the modified prediction models over real-life hospital data collected from central China in 2013-2015. To overcome the difficulty of incomplete data, we use a latent factor model to reconstruct the missing data. We experiment on a regional chronic disease of cerebral infarction. To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics. Compared to several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 94.8% with a convergence speed which is faster than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm.

Keywords : Big data analytics, Machine Learning, Healthcare.

I. INTRODUCTION

According to a report by McKinsey [1], 50% of Americans have one or more chronic diseases, and 80% of American medical care fee is spent on chronic disease treatment. With the improvement of living standards, the incidence of chronic disease is increasing. The United States has spent an average of 2.7 trillion USD annually on chronic disease treatment. This amount comprises 18% of the entire annual GDP of the United States. The healthcare problem of chronic diseases is also very important in many other countries. In China, chronic diseases are the main cause of death, according to a Chinese report on nutrition and chronic diseases in 2015, 86.6% of deaths are caused by chronic diseases. Therefore, it is essential to perform risk assessments for chronic diseases. With the growth in medical data [2], collecting electronic health records (EHR) is increasingly convenient [3]. Besides, [4] first presented a bioinspired high-performance heterogeneous vehicular telemetric paradigm, such that the collection of mobile users' health related real-time big data can be achieved with the deployment of advanced heterogeneous vehicular networks. Chen et.al. [5]– [7] proposed a healthcare system using smart clothing for sustainable health monitoring.

A. MOTIVATION

To solve these problems, it see the structured and unstructured data in healthcare field to assess the risk of disease. First, the system use Decision tree map algorithm to generate the pattern and causes of disease. It clearly shows the diseases and sub diseases. Second, by using Map Reduce algorithm for

partitioning the data such that a query will be analyzed only in a specific partition, which will increase the operational efficiency but reduce query retrieval time.

B. PROBLEM STATEMENT

- Chronic Disease Prediction by Machine Learning over Big Data from Healthcare Communities.
- The healthcare problem of chronic diseases is also very important in many other countries.
- In China, chronic diseases are the main cause of death, according to a Chinese report on nutrition and chronic diseases in 2015, 86.6% of deaths are caused by chronic diseases. Therefore, it is essential to perform risk assessments for chronic diseases.

C. OBJECTIVES OF THE PROJECT

- The analysis accuracy is reduced when the quality of medical data in incomplete. Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks.
- However, those existing work mostly considered structured data. There is no proper methods to handle semi structured and unstructured.
- The proposed system will consider both structured and unstructured data. The analysis accuracy is increased by using Machine Learning algorithm.

II. LITERATURE SURVEY

In [1], P. Groves, B. Kayyali, D. Knott, and S. V. Kuiken, based on his experience the stacks of global healthcare data are increasing exponentially. To supply high value healthcare at lower cost there is a need felt in the direction of implementing successful big data analytics pract ice to find insight from very large data sets and improving coordination, care quality and outcomes through improved analytics of healthcare big data.

In [2], M. Chen, S. Mao, and Y. Liu, he stated In this paper, we review the background and state-of-the-art of big data. We first introduce the general background of big data and review related technologies, such as could computing, Internet of Things, data centers, and Hadoop.

In [3], P. B. Jensen, L. J. Jensen, and S. Brunak, An electronic health record is a systematic collection of electronic health information about an individual patient or population.

In [4], D. Tian, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi, We propose a bio-inspired model for making handover decision in heterogeneous wireless networks. It is based on an extended attractor selection model, which is biologically inspired by the self-adaptability and robustness of cellular response to the changes in dynamic environments.

In [5] S. Bandyopadhyay, J. Wolfson, D. M. Vock, Survival prediction models most commonly use Cox Propor-tional Hazards (CPH) models, and are frequently used in medical statis-tics and clinical practice. However, such models underperform when thepredictor variables are missing. By building Bayesian networks we au-tomatically construct a model with the most important risk factors and relationships between risk factors and Bayesian networks are able to inferthe likely values of missing data.

In [6]J. Wan, S. Tang, D. Li, S. Wang, C. Liu, Nowadays the use of Big Data is increasing in biomedical and healthcare communities, accurate analysis of medical data benefits early disease detection, patient care, and community services. Incomplete medical data reduces analysis accuracy. The framework proposes Decision Tree calculation

and Naive Bayesian calculation for disease prediction. We are applying both Naïve Bayes and Decision Tree algorithm on the data set and by checking the accuracy of each algorithm we are taking the output which is more accurate.

III. PROPOSED SYSTEM

In a proposed system we can first get the large volume of a healthcare big data, then that data is considered as training data. Naive Bayes algorithm is used for the clarification of the data. Then after the clarification the hospital data similar type of data can be stored. Then CNN extract the text characteristics automatically. In that we use a CNN MDRP algorithm that uses both structured unstructured hospital data. Selecting the characteristics automatically form a large number of data. This improves the disease prediction rather than previously selected characteristics. CNN- MDRP algorithm helps to accuracy of the result of a disease prediction over a large volume of data from hospital.

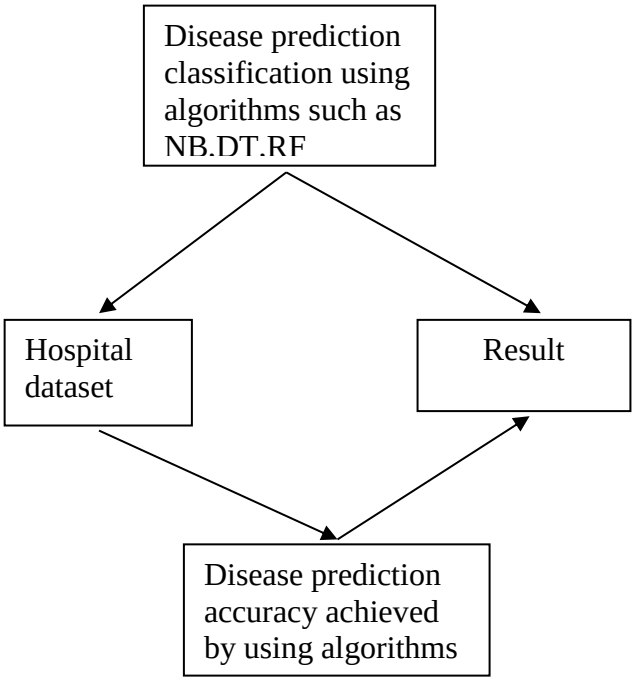

Figure 1 : Processing Steps in Disease Prediction

- We combine the structured and unstructured data in healthcare field to assess the risk of disease. First, we used latent factor model to reconstruct the missing data from the medical records collected from a hospital in central China. Second, by using statistical knowledge, we could determine the major chronic diseases in the region. Third, to handle structured data, we consult with hospital experts to extract useful features.
- The disease risk model is obtained by the combination of structured and unstructured features. Through the experiment, we draw a conclusion that the performance of CNN-MDPR is better than other existing methods.

IV. ARCHITECTURE

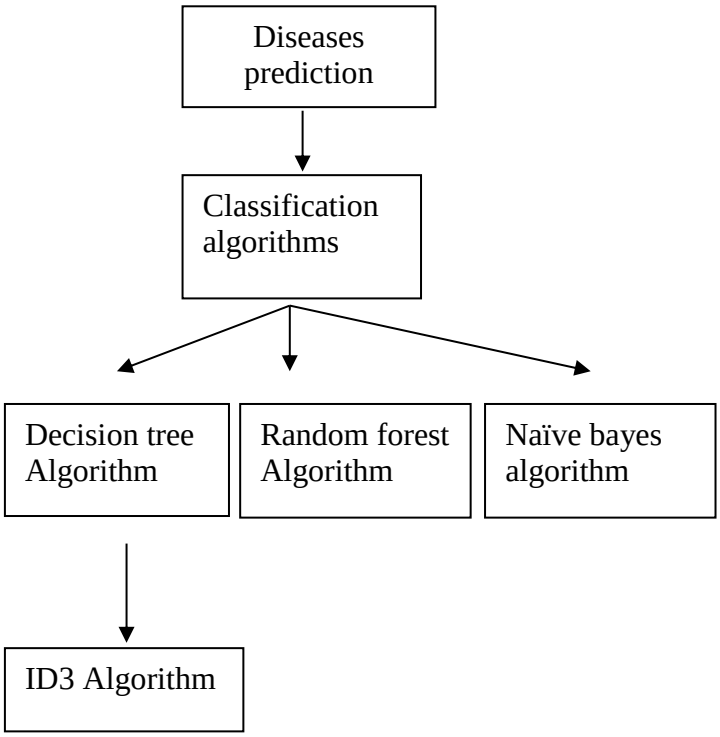

Figure 2 : Classifying of Algorithms

The above figure shows Naive Bayes classification is a simple probabilistic classifier. It requires to calculate the probability of feature attributes. In this experiment, we use conditional probability formula

to estimate discrete feature attributes and Gaussian distribution to estimate continuous feature attributes. The KNN classification is given a training data set, and the closest k instance in the training data set is found. For KNN, it is required to determine the measurement of distance and the selection of k value. In the experiment, the data is normalized at first. Then we use the Euclidean distance to measure the distance. As for the selection of parameters k, we find that the model is the best when k = 10. Thus, we choose k = 10. We choose classification and regression tree (CART) algorithm among several decision tree (DT) algorithms. To determine the best classifier and improve the accuracy of the model, the 10-fold cross-validation method is used for the training set, and data from the test set are not used in the training phase.

V. BENEFITS

- Higher Accuracy.
- We leverage not only the structureddata but also the text data of patients based on the proposedCNN-MDPR algorithm.
- We find that by combining these twodata, the accuracy rate can reach 94.80%, so as to betterevaluate the risk of cerebral infarction disease.
- To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics.

VI. RESULTS

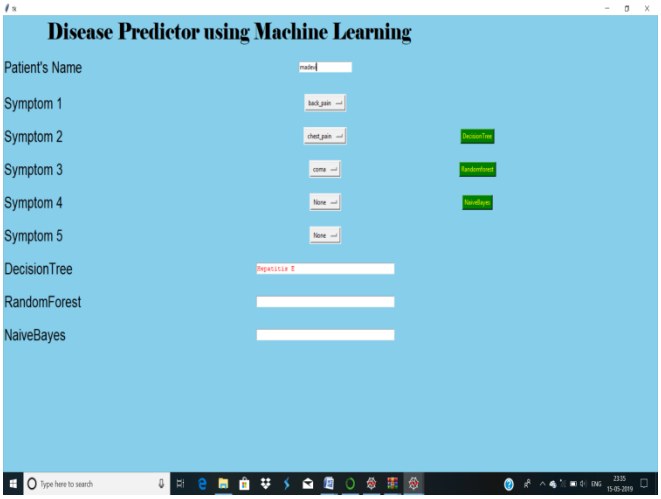

Figure 3 : Disease Prediction Using Decision Tree(DT) Algorithm

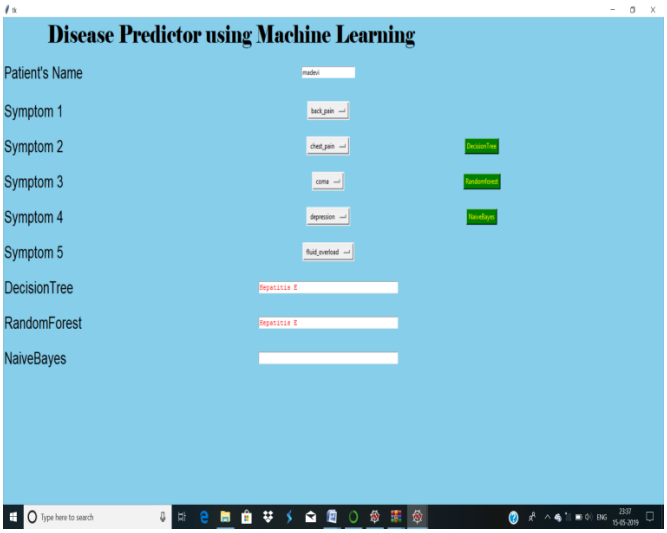

Figure 4 : Disease Prediction Using Random Forest(RF) Algorithm

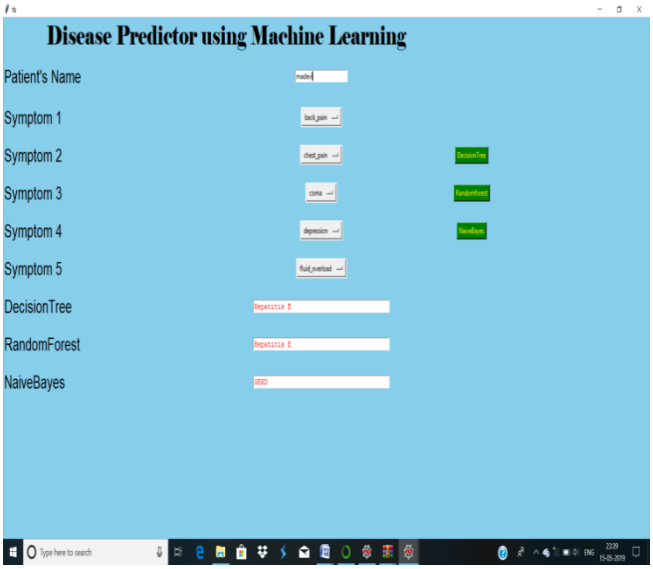

Figure 5 : Disease Prediction Using Naïve Bayes(NB) Algorithm

VII CONCLUSION

Machine learning Decision tree map algorithm by using structured and unstructured data from hospital. It also uses Map Reduce algorithm for partitioning the data. To the highest of gen, none of the current work attentive on together data types in the zone of remedial big data analytics. Compared to several typical calculating algorithms, the scheming accuracy of our proposed algorithm reaches 94.8% with an regular speed which is quicker than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm and produces report. The report consists of possibility of occurrences of diseases.

VII. REFERENCES

[1]. P. Groves, B. Kayyali, D. Knott, and S. V. Kuiken, "The 'big data' revolution in healthcare: Accelerating value and innovation," 2016.

[2]. M. Chen, S. Mao, and Y. Liu, "Big data: A survey," Mobile Networks and Applications, vol. 19, no. 2, pp. 171–209, 2014.

[3]. S. Bandyopadhyay, J. Wolfson, D. M. Vock, G. Vazquez-Benitez, G. Adomavicius, M. Elidrisi, P. E. Johnson, and P. J. O'Connor, "Data mining for censored qtime-to-event data: a bayesian network model for predicting cardiovascular risk from electronic health record data," Data Mining and Knowledge Discovery, vol. 29, no. 4, pp. 1033–1069, 2015..

[4]. D. Tian, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi, "A dynamic and self-adaptive network selection method for multimode communications in heterogeneous vehicular telematics," IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 6, pp. 3033–3049, 2015.

[5]. P. B. Jensen, L. J. Jensen, and S. Brunak.

[6]. J. Wan, S. Tang, D. Li, S. Wang, C. Liu, H.

[7]. N. Nori, H. Kashima, K. Yamashita, H. Ikai, and Y. Imanaka, "Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015, pp. 855–864.

[8]. K. Hwang, M. Chen, "Big Data Analytics for Cloud/IoT and Cognitive Computing," Wiley, U.K., ISBN: 9781119247029, 2017.

[9]. S.-M. Chu, W.-T. Shih, Y.-H. Yang, P.-C. Chen, and Y.-H. Chu, "Use of traditional chinese medicine in patients with hyperlipidemia: A population-based study in taiwan," Journal of ethnopharmacology, vol. 168, pp. 129–135, 2015.

[10]. B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," Data Mining and Knowledge Discovery, vol. 29, no. 4, pp. 1070–1093, 2015

Cite this article as : Shreekanth Jogar, Pavankumar Naik, Veeramma Vyapari, Madevi Vaddar, Kavita Dambal, Bheemavva Hatti, "Chronic Diseases Prediction over Bigdata by using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 3, pp. 246-250, May-June 2019. Available at doi : https://doi.org/10.32628/CSEIT195350
Journal URL : http://ijsrcseit.com/CSEIT195350