Disease Prediction using Big Data

Akash C. Jamgade[1] Prof. S. D. Zade[2]
[1,2]Department of Computer Science and Engineering
[1,2]Priyadarshini Institute of Engineering & Technology Nagpur, India

Abstract— Big data is growing exponentially, and along with that is the technology and new algorithms being developed. With the accumulation of big data, Machine learning and Artificial Intelligence are getting implemented into newer spheres. One of the fields in the sphere is Healthcare and Biomedical. Early disease prediction, patient care, and community services can be made possible using this accumulation of big data, with the help of Machine Learning. Though predicting diseases using Machine Learning can be implemented, but the accuracy of prediction is reduced due to incomplete medical data. Moreover different regions have different chronic diseases depending on the geographical conditions and community, an outbreak of disease. To overcome the problem of incomplete data, an approach of the latent factor model is used to reconstruct the missing data. In this paper, we propose a k-mean algorithm with the help of structured and unstructured data directly from hospital and research institutes. None of the existing work is focused on both data types in the field of medical big data analytics. Compared to several previous prediction algorithms, our prediction algorithm has the prediction accuracy of approximately 95%.

Keywords: Big Data, Healthcare, Machine Learning, K-Mean Algorithm

I. INTRODUCTION

Most of the medical care fees are spent on chronic disease treatment. The increase in chronic disease incidence has exponentially increased the data accumulation. This increase in chronic diseases has made it possible and essential to collect Electronic Health Records (EHR) and it is very convenient to do so. Besides the collection of Mobile users' health-related real-time data big data can be easily achieved with advance heterogeneous vehicular networks. But as the data accumulation in the field of medical data has made it possible to implement a system which is faster and much more accurate than that of traditional disease prediction systems. With the development of big data analytics, more and more attention can be paid on disease prediction from the perspective of the big data analysis. Many pieces of research have been already conducted by selecting various characteristics automatically from a large number of data which improves the accuracy of risk classification, rather than previously selected characteristics. However, existing research work considered structured data. Machine learning gives the methodology and different approaches which can prove to be very useful in disease prediction using machine learning. The main focus is on to use machine learning in healthcare to supplement patient care for better results. Machine learning has made easier to identify different diseases and diagnosis them correctly. Predictive analysis with the help of efficient multiple machine learning algorithms helps to predict the disease more correctly and help treat patients.

Major challenge is how to extract the information from these data because the amount is very large so some data mining and machine learning techniques can be used. Also, the expected outcome and scope of this project is that if disease can be predicted than early treatment can be given to the patients which can reduce the risk of life and save life of patients and cost to get treatment of diseases can be reduced up to some extent by early recognition. The rapid adoption of electronic health records has created a wealth of new data about patients, which is a goldmine for improving the understanding of human health. The k-mean algorithm is used to predict diseases using patient treatment history and health data.

II. EXISTING SYSTEM

Prediction using traditional disease risk model usually involves a machine learning and supervised learning algorithm which uses training data with the labels for the training of the models. High-risk and Low-risk patient classification is done in groups test sets. But these models are only valuable in clinical situations and are widely studied. A system for sustainable health monitoring using smart clothing by Chen et.al. He thoroughly studied heterogeneous systems and was able to achieve the best results for cost minimization on the tree and simple path cases for heterogeneous systems.

The information of patient's statistics, test results, and disease history is recorded in EHR which enables to identify potential data-centric solutions which reduce the cost of medical case studies. Bates et al. propose six applications of big data in the healthcare field. Existing systems can predict the diseases but not the subtype of diseases. It fails to predict the condition of people.

The predictions of diseases have been non-specific and indefinite

III. PROPOSED SYSTEM

In this paper, we have combined the structure and unstructured data in healthcare fields that let us assess the risk of disease. The approach of the latent factor model for reconstructing the missing data in medical records which are collected from the hospital. And by using statistical knowledge, we could determine the major chronic diseases in a particular region and in particular community. To handle structured data, we consult hospital experts to know useful features.

In the case of unstructured text data, we select the features automatically with the help of k-mean algorithm. We propose a k-mean algorithm for both structured and unstructured data.

IV. SYSTEM ARCHITECTURE

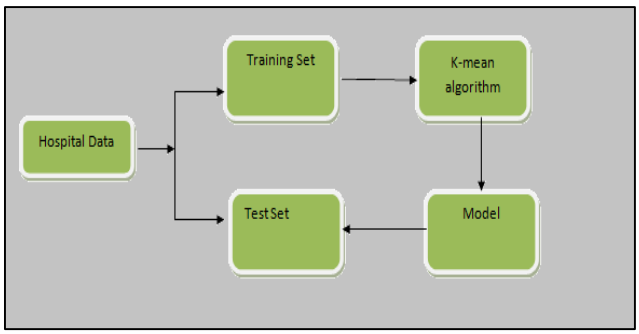

Fig. 1: System Architecture

The disease risk model is obtained by combining both structured and unstructured features.

V. CONCLUSIONS

With the proposed system, higher accuracy can be achieved. We not only use structured data, but also the text data of the patient based on the proposed k-mean algorithm. To find that out, we combine both data, and the accuracy rate can be reached up to 95%. None of the existing system and work is focused on using both the data types in the field of medical big data analytics. We propose a K-Mean clustering algorithm for both structured and unstructured data. The disease risk model is obtained by combining both structured and unstructured features.

ACKNOWLEDGMENT

I express my sincere gratitude towards my guide of Prof. S. D. Zade for their constant help, encouragement and inspiration throughout the project work.

Also I would like to thank the Head of Computer Science and Engineering Department Dr. P. S. Prasad for him valuable guidance, ability to motive me and even willingness to solve difficulty made it possible to make my project unique and made task easier.

My sincere thanks to Principal, Dr. V. M. Nanoti for providing me necessary facility to carry out the work.

I take this opportunity to express my hearty thanks to all those who help me in the completion of my project work. I am very grateful to authors of various articles and journals, for helping me become aware of the current ongoing research in the related field.

Last, but not the least, I would like to thank my classmates for their valuable comments, suggestions and unconditional support.

REFERENCES

[1] P. Groves, B. Kayyali, D. Knott, and S. V. Kuiken, "The'bigdata'revolution in healthcare: Accelerating value and innovation, big-data revolution".
[2] J. Wang, M. Qiu, and B. Guo, "Enabling real-time information service on telehealth system over cloud-based big data platform," Journal of Systems Architecture, vol. 72, pp. 69–79, 2017
[3] D. Oliver, F. Daly, F. C. Martin, and M. E. McMurdo, "Risk factors and risk assessment tools for falls in hospital in-patients: a systematic review," Age and ageing, vol. 33, no. 2, pp. 122–130, 2004
[4] S. Marcoon, A. M. Chang, B. Lee, R. Salhi, and J. E. Hollander, "Heart score to further risk stratify patients with low timi scores," Critical pathways in cardiology, vol. 12, no. 1, pp. 1–5, 2013.
[5] S. Bandyopadhyay, J. Wolfson, D. M. Vock, G. Vazquez-Benitez, G. Adomavicius, M. Elidrisi, P. E. Johnson, and P. J. O'Connor, "Datamining for censored time-to-event data: a bayesian network model for predicting cardiovascular risk from electronic health record data,"
[6] Data Mining and Knowledge Discovery, vol. 29, no. 4, pp. 1033–1069, 2015.

