ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 12, December 2017

A Survey on Disease Prediction by Machine Learning Over Big Data from Healthcare Communities

Santosh S Shinde, Prof. Shweta Joshi

Department of CE, Flora Institute of Technology, Pune, India

HOD, Department of CE, Flora Institute of Technology, Pune, India

ABSTRACT: The fast growing sector of big data analytics has started to play a key role in the evolution of healthcare practices and research. It has provided tools for gathering, managing, analyzing and assimilating large, structured and unstructured volumes of data produced by current healthcare systems. Big Data Analysis has recently been applied to help the process of delivering attention and exploration of diseases. However, the rate of adoption and development of research in this space is still hampered by some fundamental problems inherent in the big data paradigm. In this paper, we have optimized automatic learning algorithms for the effective prediction of chronic disease epidemics in frequent disease communities. We propose a new convolutional neural networks (CNN-MDRP) multimodal disease prediction algorithm using structured and unstructured hospital data.

KEYWORDS*:* Big data analytics; Machine Learning; Healthcare

I. INTRODUCTION

The concept of big data is not new; however the way it is defined is constantly changing. Various attempts at defining big data essentially characterize it as a collection of data elements whose size, speed, type, and/or complexity require one to seek, adopt, and invent new hardware and software mechanisms for archiving, analyzing and displaying data successfully. Healthcare is a prime example of how the three V's of data first is velocity, second is variety, and third one is volume are an innate aspect of the data it produces. This data is spread among multiple healthcare systems, health insurers, researchers, government entities, and so forth. Furthermore, each of these data repositories is siloed and inherently incapable of providing a platform for global data transparency. To add to the three V's, the veracity of healthcare data is also critical for its meaningful use towards developing translational research.

With the development of big data technology, more attention has been paid to disease prediction from the perspective of big data analysis; various researches have been conducted by selecting the characteristics automatically from a large number of data to improve the accuracy of risk classification rather than the previously selected characteristics. However, those existing work mostly considered structured data. For unstructured data, for example, using convolutional neural network (CNN) to extract text characteristics automatically has already attracted wide attention and also achieved very good results. However, to the best of our knowledge, none of previous work handles medical text data by CNN. Furthermore, there is a large difference between diseases in different regions, primarily because of the diverse climate and living habits in the region. Thus, risk classification based on big data analysis, the following challenges remain: How should the missing data be addressed? How should the main chronic diseases in a certain region and the main characteristics of the disease in the region be determined? How can big data analysis technology be used to analyze the disease and create a better model? To solve these problems, we combine the structured and unstructured data in healthcare field to assess the risk of disease.

Background:

1. Only work on structured data.
2. CNN –UDRP algorithm used for structured data.

DOI: 10.15680/IJIRCCE.2017. 0512051

Motivation:

1. Medical datasets are often not balanced in their class labels.
2. Most of the existing classification methods tend to perform poorly on a dataset which is extremely imbalanced.
3. In Previous only work on structured Data.

Goals/Objectives:

1. Disease Prediction for Structured Data we use machine learning algorithms i.e., Naive Bayesian (NB), K-nearest Neighbors (KNN), and Decision Tree (DT) algorithms.
2. Disease Prediction accuracy is calculated using CNN-MDRP algorithm for structured or Unstructured Data.

II. RELATED WORK

1]**P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care:**Clinical data describing the phenotypes and treatment of patients represents an underused data source that has much greater research potential than is currently realized. Mining of electronic health records (EHRs) has the potential for establishing new patient-stratification principles and for revealing unknown disease correlations. Integrating EHR data with genetic data will also give a finer understanding of genotype-phenotype relationships. However, a broad range of ethical, legal and technical reasons currently hinder the systematic deposition of these data in EHRs and their mining. Here, we consider the potential for furthering medical research and clinical care using EHR data and the challenges that must be overcome before this is a reality.

2]**M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, C. Youn, "Wearable 2.0: Enable Human-Cloud Integration in Next Generation Healthcare System,"** ***IEEECommunications*****, Vol. 55, No. 1, pp. 54–61, Jan. 2017:**With the rapid development of the Internet of Things, cloud computing, and big data, more comprehensive and powerful applications become available. Meanwhile, people pay more attention to higher QoE and QoS in a źterminal- cloudź integrated system. Specifically, both advanced terminal technologies (e.g., smart clothing) and advanced cloud technologies (e.g., big data analytics and cognitive computing in clouds) are expected to provide people with more reliable and intelligent services. Therefore, in this article we propose a Wearable 2.0 healthcare system to improve QoE and QoS of the next generation healthcare system. In the proposed system, washable smart clothing, which consists of sensors, electrodes, and wires, is the critical component to collect users physiological data and receive the analysis results of users health and emotional status provided by cloud-based machine intelligence.

3] M. Chen, Y. Ma, J. Song, C. Lai, B. Hu, "Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring,"ACM/Springer Mobile Networks and Applications' Vol. 21, No. 5, pp.825C845, 2016.: Traditional wearable devices have various shortcomings, such as comfortableness for long-term wearing, and insufficient accuracy, etc. Thus, health monitoring through traditional wearable devices is hard to be sustainable. In order to obtain healthcare big data by sustainable health monitoring, we design "Smart Clothing", facilitating unobtrusive collection of various physiological indicators of human body. To provide pervasive intelligence for smart clothing system, mobile healthcare cloud platform is constructed by the use of mobile internet, cloud computing and big data analytics. This paper introduces design details, key technologies and practical implementation methods of smart clothing system. Typical applications powered by smart clothing and big data clouds are presented, such as medical emergency response, emotion care, disease diagnosis, and real-time tactile interaction. Especially, electrocardiograph signals collected by smart clothing are used for mood monitoring and emotion detection. Finally, we highlight some of the design challenges and open issues that still need to be addressed to make smart clothing ubiquitous for a wide range of applications.

4] J. Wang, M. Qiu, and B. Guo, "Enabling real-time information service on telehealth system over cloud-based big data platform," ***Journal of SystemsArchitecture*****, vol. 72, pp. 69–79, 2017:**We design a data coherence protocol for the PHR-based distributed system.We propose a flow estimating algorithm for the telehealth cloud system.We apply several predicting methods for the future bandwidth consumption.We present a telehealth framework for bandwidth balance on emergency. A telehealth system covers both clinical and nonclinical uses, which not only provides store-and-forward data services to be offline studied by relevant specialists, but also monitors the real-time physiological data through ubiquitous sensors to support remote telemedicine. However, the current telehealth systems do not consider the velocity and veracity of the big-data system in the medical context. Emergency events generate a large amount of the real-time data, which should be stored in the data center, and forwarded to remote hospitals. Furthermore, patients' information is scattered on the distributed data center, which cannot provide a high-efficient remote real-time service. In this paper, we proposes a probability-based bandwidth model in a telehealth cloud system, which helps cloud broker to provide a high performance allocation of computing nodes and links. This brokering mechanism considers the location protocol of Personal Health Record (PHR) in cloud and schedules the real-time signals with a low information transfer between different hosts. The broker uses several bandwidth evaluating methods to predict the near future usage of bandwidth in a telehealth context. The simulation results show that our model is effective at determining the best performing service, and the inserted service validates the utility of our approach.

5] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, “Big data in health care: using analytics to identify and manage high-risk and high-cost patients,” *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014:The US health care system is rapidly adopting electronic health records, which will dramatically increase the quantity of clinical data that are available electronically. Simultaneously, rapid progress has been made in clinical analytics-techniques for analyzing large quantities of data and gleaning new insights from that analysis-which is part of what is known as big data. As a result, there are unprecedented opportunities to use big data to reduce the costs of health care in the United States. We present six use cases-that is, key examples-where some of the clearest opportunities exist to reduce costs through the use of big data: high-cost patients, readmissions, triage, decompensation (when a patient's condition worsens), adverse events, and treatment optimization for diseases affecting multiple organ systems. We discuss the types of insights that are likely to emerge from clinical analytics, the types of data needed to obtain such insights, and the infrastructure-analytics, algorithms, registries, assessment scores, monitoring devices, and so forth-that organizations will need to perform the necessary analyses and to implement changes that will improve care while reducing costs. Our findings have policy implications for regulatory oversight, ways to address privacy concerns, and the support of research on analytics.

6]**Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, “Healthcps: Healthcare cyber-physical system assisted by cloud and big data:**The advances in information technology have witnessed great progress on healthcare technologies in various domains nowadays. However, these new technologies have also made healthcare data not only much bigger but also much more difficult to handle and process. Moreover, because the data are created from a variety of devices within a short time span, the characteristics of these data are that they are stored in different formats and created quickly, which can, to a large extent, be regarded as a big data problem. To provide a more convenient service and environment of healthcare, this paper proposes a cyber-physical system for patient-centric healthcare applications and services, called Health-CPS, built on cloud and big data analytics technologies. This system consists of a data collection layer with a unified standard, a data management layer for distributed storage and parallel computing, and a data-oriented service layer. The results of this study show that the technologies of cloud and big data can be used to enhance the performance of the healthcare system so that humans can then enjoy various smart healthcare applications and services.

8] K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, “Localization based on social big data analysis in the vehicular networks,” *IEEETransactions on Industrial Informatics*, 2016: Location-based services, especially for vehicular localization, are an indispensable component of most technologies and applications related to the vehicular networks. However, because of the randomness of the vehicle movement and the complexity of a driving environment, attempts to develop an effective localization solution face certain difficulties. In this paper, an overlapping and hierarchical social clustering model (OHSC) is first designed to classify the vehicles into different social clusters by exploring the social relationship between them. By using the results of the OHSC model, we propose a social-based localization algorithm (SBL) that use location prediction to assist in global localization in the vehicular networks. The experiment results validate the performance of the OHSC model and show that the presented SBL algorithm demonstrates superior localization performance compared with the existing methods.

9] 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:A small number of significant falls risk factors emerged consistently, despite the heterogeneity of settings namely gait instability, agitated confusion, urinary incontinence/frequency, falls history and prescription of 'culprit' drugs (especially sedative/hypnotics). Simple risk assessment tools constructed of similar variables have been shown to predict falls with sensitivity and specificity in excess of 70%, although validation in a variety of settings and in routine clinical use is lacking. Effective falls interventions in this population may require the use of better-validated risk assessment tools, or alternatively, attention to common reversible falls risk factors in all patients.

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 andKnowledge Discovery*, vol. 29, no. 4, pp. 1070–1093, 2015: This paper investigates the patient risk prediction problem in the context of active learning with relative similarities. Active learning has been extensively studied and successfully applied to solve real problems. The typical setting of active learning methods is to query absolute questions. In a medical application where the goal is to predict the risk of patients on certain disease using Electronic Health Records (EHR), the absolute questions take the form of "Will this patient suffer from Alzheimer's later in his/her life?", or "Are these two patients similar or not?". Due to the excessive requirements of domain knowledge, such absolute questions are usually difficult to answer, even for experienced medical experts. In addition, the performance of absolute question focused active learning methods is less stable, since incorrect answers often occur which can be detrimental to the risk prediction model. In this paper, alternatively, we focus on designing relative questions that can be easily answered by domain experts. The proposed relative queries take the form of "Is patient A or patient B more similar to patient C?", which can be answered by medical experts with more confidence. These questions poll relative information as opposed to absolute information, and even can be answered by non-experts in some cases. In this paper we propose an interactive patient risk prediction method, which actively queries medical experts with the relative similarity of patients. We explore our method on both benchmark and real clinic datasets, and make several interesting discoveries including that querying relative similarities is effective in patient risk prediction, and sometimes can even yield better prediction accuracy than asking for absolute questions.

11] S. Zhai, K.-h. Chang, R. Zhang, and Z. M. Zhang, "Deepintent: Learning attentions for online advertising with recurrent neural networks : In this paper, we investigate the use of recurrent neural networks (RNNs) in the context of search-based online advertising. We use RNNs to map both queries and ads to real valued vectors, with which the relevance of a given (query, ad) pair can be easily computed. On top of the RNN, we propose a novel attention network, which learns to assign attention scores to different word locations according to their intent importance (hence the name DeepIntent). The vector output of a sequence is thus computed by a weighted sum of the hidden states of the RNN at each word according their attention scores. We perform end-to-end training of both the RNN and attention network under the guidance of user click logs, which are sampled from a commercial search engine. We show that in most cases the attention network improves the quality of learned vector representations, evaluated by AUC on a manually labeled dataset. Moreover, we highlight the effectiveness of the learned attention scores from two aspects: query rewriting and a modified BM25 metric. We show that using the learned attention scores, one is able to produce sub-queries that are of better qualities than those of the state-of-the-art methods. Also, by modifying the term frequency with the attention scores in a standard BM25 formula, one is able to improve its performance evaluated by AUC.

III. PROPOSED SYSTEM APPROACH

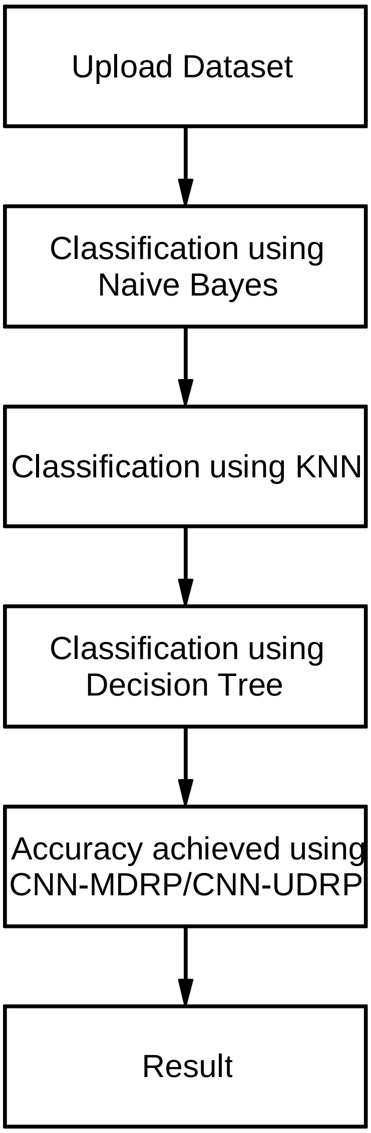
Upload Dataset
Classification using Naive Bayes
Classification using KNN
Classification using Decision Tree
Accuracy achieved using CNN-MDRP/CNN-UDRP
Result

Fig 1. Proposed System Architecture

In Proposed Work, we propose a new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital. To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data Analytics. With the development of big data analytics technology, more attention has been paid to disease prediction from the perspective of big data analysis, various researches have been conducted by selecting the characteristics automatically from a large number of data to improve the accuracy of risk classification, rather than the previously selected characteristics. However, those existing work mostly considered structured data. For unstructured data, for example, using convolutional neural network (CNN) to extract text characteristics automatically has already attracted wide attention and also achieve

IV. CONCLUSION

In this paper, we propose a machine learning and new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital. In existing work not focused on both data types in the area of healthcare 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 (CNNUDRP) algorithm.

REFERENCES

1] P. Groves, B. Kayyali, D. Knott, and S. V. Kuiken, “The’big data’revolution in healthcare: Accelerating value and innovation,big-data revolution”.

2] P. B. Jensen, L. J. Jensen, and S. Brunak, “Mining electronic health records: towards better research applications and clinical care

3] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, C. Youn, “Wearable 2.0: Enable Human-Cloud Integration in Next Generation Healthcare System,” *IEEECommunications*, Vol. 55, No. 1, pp. 54–61, Jan. 2017

4] M. Chen, Y. Ma, J. Song, C. Lai, B. Hu, ”Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring,”ACM/Springer Mobile Networks and Applications’ Vol. 21, No. 5, pp.825C845, 2016

5] J. Wang, M. Qiu, and B. Guo, “Enabling real-time information service on telehealth system over cloud-based big data platform,” *Journal of SystemsArchitecture*, vol. 72, pp. 69–79, 2017

6] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, “Big data in health care: using analytics to identify and manage high-risk and high-cost patients,” *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014

7] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, “Healthcps: Healthcare cyber-physical system assisted by cloud and big data

8] K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, “Localization based on social big data analysis in the vehicular networks,” *IEEETransactions on Industrial Informatics*, 2016

9] 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

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 andKnowledge Discovery*, vol. 29, no. 4, pp. 1070–1093, 2015

11] S. Zhai, K.-h. Chang, R. Zhang, and Z. M. Zhang, “Deepintent: Learning attentions for online advertising with recurrent neural networks