International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 3
ISSN: 2454-4248
400 – 405

Prediction of Disease Using Machine Learning over Big Data-Survey

Lohith S Y
Department Of Studies in Computer Science & Engineering ,
University BDT College Of Engineering
(A Constituent College Of VTU,Belagavi),
Davanagere,Karnataka

Dr. Mohamed Rafi
Professor
Department Of Computer Science & Engineering,
University BDT College Of Engineering
(A Constituent College Of VTU,Belagavi),
Davanagere,Karnataka

Abstract— With massive information development in medical specialty and aid community, precise analysis of medical information advantages premature disease detection, patient care and community services. although, the analysis accuracy is reduced once the standard of medical information is incomplete. moreover, completely different regions exhibit distinctive characteristics of bound regional diseases, which can weaken the prediction of illness outbreaks. during this paper, we tend to contour machine learning algorithms for effective prediction of chronic malady eruption in disease-frequent communities. we tend to experiment the tailored prediction models over real-life hospital information collected from central China in 2013-2015. to beat the problem of incomplete information, we tend to use a latent issue model to build the missing information. we tend to experiment on a regional chronic illness of cerebral infarction. we tend to propose a replacement convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithmic program victimisation structured and unstructured information from hospital. To the simplest of our data, none of the prevailing work targeted on each information varieties within the space of medical massive information analytics. Compared to many typical prediction algorithms, the prediction accuracy of our projected algorithmic program reaches ninety four.8% with a convergence speed that is faster than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithmic program.

Key words: *Data Mining ,Prediction, Risk Factors.*

******

I. INTRODUCTION

According to a report by McKinsey, 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 prblem 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, collecting electronic health records (EHR) is increasingly convenient. Besides, first presented a bio
inspired 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. proposed a healthcare system using smart clothing for sustainable health monitoring. had thoroughly studied the heterogeneous systems and achieved the best results for cost minimization on tree and simple path cases for heterogeneous systems. Patients' statistical information, test results and disease history are recorded in the EHR, enabling us to identify potential data-centric solutions to reduce the costs of medical case studies. proposed an efficient flow estimating algorithm for the telehealth cloud system and designed a data coherence protocol for the PHR(Personal Health Record)-based distributed system. Bates et al proposed six applications of big data in the field of healthcare. Qiu et al. proposed an optimal big data sharing algorithm to handle the complicate data set in tele health with cloud techniques. One of the applications is to identify high-risk patients which can be utilized to reduce medical cost since high-risk patients often require expensive healthcare. Moreover, in the first paper proposing healthcare cyber-physical system, it innovatively brought forward the concept of prediction-based healthcare applications, including health risk assessment. Prediction using traditional disease risk models usually involves a machine learning algorithm (e.g., logistic regression and regression analysis, etc.), and especially a supervised learning algorithm by the use of training data with labels to train the model. In the test set, patients can be classified into groups of either high-risk or low-risk. These models are valuable in clinical situations and are widely studied. However, these schemes have the following characteristics and defects. The data set is typically small, for patients and diseases with specific conditions, the characteristics are selected through experience. However, these pre-selected characteristics maybe not satisfy the changes in the disease and its influencing factors.

 ISSN: 2454-4248
400 – 405

Review of Literature: 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 achieved very good results. However, to the best of our knowledge, none of previous work handle Chinese 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. 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. For unstructured text data, we select the features automatically using CNN algorithm. Finally, we propose a novel CNN-based multimodal disease risk prediction (CNN-MDRP) algorithm for structured and unstructured data. 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.

David W. Bates, Suchi Saria, Lucila Ohno-Machado et al proposed a Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. The US health care system is hurriedly adopting electronic health records, which will considerably increase the amount of clinical data that are available electronically. at the same time, rapid progress has been made in clinical analytics—techniques for analyzing large quantity 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 opportunity 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, decomposition (when a patient's condition worsens), adverse events, and treatment optimization for diseases disturbing 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, judgment 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. The choice of these specific use cases that we have Discussed in this article can be debated. None the less, we believe that they will be among those that deliver the greatest value for health care organizations in the near term.

Yin Zhang, Meikang Qiu, Chun-Wei Tsai, Mohammad et al proposed a Health-CPS: 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. a smart health system assisted by cloud and big data, which includes 1) a unified data collection layer for the integration of public medical resources and personal health devices, 2) a cloud-enabled and data-driven platform for multisource heterogeneous healthcare data storage and analysis, and 3) a unified API for developers and a unified interface for users.

Sunayan Bandyopadhyay,Julian Wolfson, David M. Vock et al proposed a Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data. Models for predicting the risk of cardiovascular (CV) events based on individual patient characteristics are important tools for managing patient care. Most current and commonly used risk predictionmodels have been built from carefully selected epidemiological cohorts. However, the homogeneity and limited size of such cohorts restrict the predictive power and eneralizability of these risk models to other populations. Electronic health data (EHD) from large health care systems provide access to data on large, heterogeneous, and contemporaneous patient populations. The

 ISSN: 2454-4248
400 – 405

unique features and challenges of EHD, including missing risk factor information, non-linear relationships between risk factors and CV event outcomes, and differing effects from different patient subgroups, demand novel machine learning approaches to risk model development. In this paper, we present a machine learning approach based on Bayesian networks trained onEHDto predict the probability of having aCVevent within 5 years. In such data, event status may be unknown for some individuals, as the event time is right-censored due to disenrollment and incomplete follow-up. Since many traditional dataminingmethods are not well-suited for such data, we describe howto modify both modeling and assessment techniques to account for censored observation times. We show that our approach can lead to better predictive performance than the Cox proportional hazards model (i.e., a regression-based approach commonly used for censored, time-to-event data) or a Bayesian network with ad hoc approaches to right-censoring. Our techniques are motivated by and illustrated on data from a large US Midwestern health care system.

Buyue Qian, Xiang Wang, Nan Cao et al proposed a A relative similarity based method for interactive patient risk prediction. 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.

Keke Gai, Meikang Qiu et al proposed a Optimal Big Data Sharing Approach for Tele-health in Cloud Computing. The rapid development of tele-health systems have received driving engagements from various emerging techniques, such as big data and cloud computing. Sharing data among multiple tele-health systems is an adaptive approach for improving service quality via the network-based technologies. However, current implementations of data sharing in cloud computing is still facing the restrictions caused by the networking capacities and virtual machine switches. In this paper, we focus on the problem of data sharing obstacles in cloud computing and propose an approach that uses dynamic programming to produce optimal solutions to data sharing mechanisms. The proposed approach is called Optimal Telehealth Data Sharing Model (OTDSM), which considers transmission probabilities, maximizing network capacities, and timing constraints. Our experimental results have proved the flexibility and adoptability of the proposed method. An approach of creating optimal solutions to big data sharing in cloud-based tele-health systems. We considered three vital factors that highly impacted on the real-time tele-health services, including the transmission time, network capacity, and transmission success probability.

Kai Lin, Jiming Luo, Long Hu et al proposed Localization based on Social Big Data Analysis in the Vehicular Networks . Location Based Services (LBS), 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 firstly 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.

Min Chen, Yujun Ma, Jeungeun Song et al proposed Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring. Traditional wearable devices have various shortcomings, such as uncomfortableness 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

 ISSN: 2454-4248
400 – 405

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.

Kai Lin, Min Chen, Jing Deng, Mohammad Mehedi Hassan et al proposed Enhanced Fingerprinting and Trajectory Prediction for IoT Localization in Smart Buildings. Location service is one of the primary services in smart automated systems of Internet of Things (IoT). For various location-based services, accurate localization has become a key issue. Recently, research on IoT localization systems for smart buildings has been attracting increasing attention. In this paper, we propose a novel localization approach that utilizes the neighbor relative received signal strength to build the fingerprint database and adopts a Markov-chain prediction model to assist positioning. The approach is called the novel localization method (LNM) in short. In the proposed LNM scheme, the history data of the pedestrian's locations are analyzed to further lower the unpredictable signal fluctuations in a smart building environment, meanwhile enabling calibration-free positioning for various devices. The performance evaluation conducted in a realistic environment shows that the presented method demonstrates superior localization performance compared with well-known existing schemes, especially when the problems of device heterogeneity and WiFi signals fluctuation exist.

Jihe Wang, Meikang Qiu et al proposed Enabling Real-time Information Service on Telehealth System over Cloud-based Big Data Platform. 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 don't 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.

Min Chen, Yujun Ma, Yong Li, Di Wu et al proposed Wearable 2.0: Enabling Human-Cloud Integration in Next Generation Healthcare Systems. 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 "terminal-cloud" 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.

II. SUMMARY OF LITERATURE SURVEY:

We design a distributed data managing framework for telehealth system, which includes BSN, cloud system, and remote hospital end. By analyzing the features of data processing with medical applications, we provide a decentralized data coherence protocol to solve the performance problems by current design. Our model measures the bandwidth consumption between any node pair in cloud so that the bandwidth can be calculated in each interval. The experimental results show that the bandwidth predicting error is limited in 10%, which provides cloud with a flexible methods (4 types of predicting algorithms) to estimate the bandwidth resources to nodes. Furthermore, a case study shows that our method is able to support finding the most appropriate bandwidth-estimating algorithm for underlining telehealth applications. In future, we plan to apply our approach to many real-world projects and get the feedback from the collaboration with hospitals. Furthermore, we will work on this topic with more advanced approach, such as hidden Markov model to enhance the performance bandwidth estimating.

III. PROPOSED-SYSTEM:

For disease risk modelling, the accuracy of risk prediction depends on the diversity feature of the hospital data, i.e., the better is the feature description of the disease, the higher the accuracy will be. For some simple disease, e.g., hyper lipidemia, only a few features of structured data can get a good description of the disease, resulting in fairly good effect of disease risk prediction. But for a complex disease, such as cerebral infarction mentioned in the paper, only using features

 ISSN: 2454-4248
400 – 405

of structured data is not a good way to describe the disease. As seen from The corresponding accuracy is low, which is roughly around 50%. Therefore, in this paper, we leverage not only the structured data but also the text data of patients based on the proposed CNN-MDPR algorithm. We find that by combining these two data, the accuracy rate can reach 94.80%, so as to better evaluate the risk of cerebral infarction disease.

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. 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.

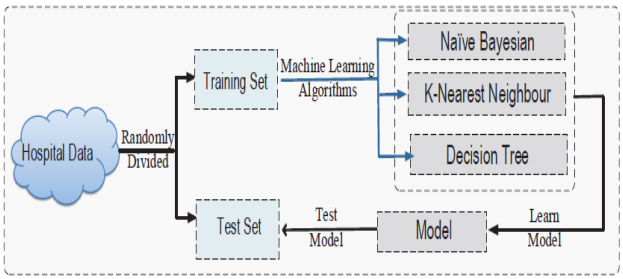

Fig. . The three machine learning algorithms userd in our disease prediction experiments.

IV. MODULES:

For dataset, according to the different characteristics of the patient and the discussion with doctors, we will focus on the following three datasets to reach a conclusion.

- ✓ Structured data (S-data): use the patient's structured data to predict whether the patient is at high-risk of cerebral infarction.
- ✓ Text data (T-data): use the patient's unstructured text data to predict whether the patient is at high-risk of cerebral infarction.
- ✓ Structured and text data (S&T-data): use the S-data and T-data above to multi-dimensionally fuse the structured data and unstructured text data to predict whether the patient is at high-risk of cerebral infarction.

V. CONCLUSION:

In this paper, we tend to propose a replacement convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithmic rule exploitation structured and unstructured information from hospital. To the most effective of our data, none of the prevailing work centered on each information sorts within the space of medical massive information analytics. Compared to many typical prediction algorithms, the prediction accuracy of our projected algorithmic rule reaches 94.8% with a convergence speed that is quicker than that of the CNN-based unimodal disease risk prediction (CNNUDRP) algorithmic rule.

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]. P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," Nature Reviews Genetics, vol. 13, no. 6, pp. 395–405, 2012.

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

[6]. 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.

[7]. M. Chen, P. Zhou, G. Fortino, "Emotion Communication System," IEEE Access, DOI: 10.1109/ACCESS.2016.2641480, 2016.

[8]. M. Qiu and E. H.-M. Sha, "Cost minimization while satisfying hard/soft timing constraints for heterogeneous embedded systems," ACM Transactions on Design Automation of Electronic Systems (TODAES), vol. 14, no. 2, p. 25, 2009.

[9]. 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.

[10]. 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.

[11]. L. Qiu, K. Gai, and M. Qiu, "Optimal big data sharing approach for tele-health in cloud computing," in Smart Cloud (SmartCloud), IEEE International Conference on. IEEE, 2016, pp. 184–189.

[12]. Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "Healthcps: Healthcare cyber-physical system assisted by cloud and big data," IEEE Systems Journal, 2015.

[13]. K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, "Localization based on social big data analysis in the

ISSN: 2454-4248
400 – 405

vehicular networks," IEEE Transactions on Industrial Informatics, 2016.

[14]. K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, "Enhanced fingerprinting and trajectory prediction for iot localization in smart buildings," IEEE Transactions on Automation Science and Engineering,vol. 13, no. 3, pp. 1294–1307, 2016.

[15]. 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.

[16]. 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.

[17]. 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 time-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.

[18]. 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.