

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

Healthcare Prediction Analysis Using Machine Learning Over Big Data

Santosh Suresh Shinde, Prof.Nita Dimble

Dept. of Computer Engineering, Flora Institute of Technology Pune, India

Asst. Professor Dept. of Computer Engineering, Flora Institute of Technology Pune, India

ABSTRACT: The Big data analytic's plays a crucial role in the progress of healthcare research.

It has provides the tools to Arrange, collect, analyze, and intake large volumes of constructed or structured, and unconstructed or unstructured data produced by current wellness care systems. Analytics of Big data has newly applied technique on backup the litigate of delivery and disease prediction. However, the acceptance rate and development of research in this space is still threaded by some fundamental problems inherent within big data patterns.

We optimized algorithms of machine learning for the more effective predict the outbreak of disease in frequent disease communities. We optimized machine learning algorithms for the more effective prevision of the eruption of disease in often disease communities. We have proposed latest multimodal disease risk prevision algorithm based on convolution neural networks (CNN-MDRP) by using unstructured and structured hospital's data.

KEYWORDS: Data Analytics; Machine Learning; Healthcare Data, Health examination, early Warning, Classification Techniques, Medical Expenses,

I. INTRODUCTION

The big data concept is not new however the way of expressing is continuously changing. Number of trial at defining big data fundamentally characterize it as a collection of data elements whose magnitude, type, speed, and intricacy require one to adopt, seek, and device new hardware and software reaction for archiving, analyzing and displaying data successfully.

Healthcare is a basic instance of how the three V's of data work first is velocity, second is variety, and third is volume are an inborn aspect of the data it raise. This data is circulating among various multiple healthcare systems, researchers, health insurers, government entities, and so forth. Furthermore, everyone of these data monument is soloed and constitutive in noncompeting of procuring a platform for transparency of global data.

To add the three V's, healthcare data's veracity is also decisive for its meaningful use towards under development translational research. With the analysis of big data technology, more attention was paid to the disease prediction from the point of view of big data analysis; various studies were performed by selecting the features automatically from a large amount of data to improve accuracy of the risk classification rather than previously selected features. However, these existing works have mainly considered structured data. For unstructured data, for example, use of the convolution neural network (CNN) to automatically extract the functionality of the text has already attracted great attention and has also achieved very good results. However, as far we know, none of the previous work has managed data from the CNN medical text. Furthermore, there is a big difference between in different regions diseases, mainly due to the diversity of climate and life habits in the region. Therefore, the classification of risks based on the analysis of big data continues to be the some following challenge: How missing data should be processed? How should main chronic diseases in given

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

region and the main characteristics of the disease in the region be determined? How can big data analytics be used to analyze the disease and how can create a better model? To solve these problems, we combine structured and unstructured data in the field of health to assess the risk of disease.

OBJECTIVES

1. To predict disease for integrated Data using machine learning algorithm. I.e. Decision Tree
2. To calculate disease prediction accuracy using CNN-MDRP algorithm for structured or Unstructured Data.

Software Requirement Specification

- **Software Environment:**
 - a. Operating System : Windows XP
 - b. Programming Language : JAVA

- **Hardware Environment:**
 - c. Processor - Pentium –IV
 - d. Speed - 1.1 GHz
 - e. RAM - 256 MB(min)
 - f. Hard Disk - 20 GB
 - g. Key Board - Standard Windows Keyboard
 - h. Mouse - Two or Three Button Mouse
 - i. Monitor - SVGA

External Interface Requirements

User Interfaces

- **Front-End software:**
 - i. Eclipse IDE
 - ii. Java

- **Back-End software:**
 - i. MongoDB Database
 - ii. Apache tomcat server

Mathematical Model and Design:

Let us consider S as a system for automatically recommends places.

$S = \{ \dots \}$

INPUT:

- Identify the inputs

$F = \{f_1, f_2, f_3, \dots, f_n\}$ 'F' as set of functions to execute commands. }

$I = \{i_1, i_2, i_3, \dots\}$ 'I' sets of inputs to the function set }

$O = \{o_1, o_2, o_3, \dots\}$ 'O' Set of outputs from the function sets, }

$S = \{I, F, O\}$

I = {Query submitted by the user, i.e. disease or symptoms }

O = {Output of desired query, i.e. symptoms or disease and accuracy }

F = {Functions implemented to get the output, i.e. machine learning algorithm and CNN-MDRP }

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

II. REVIEW OF LITERATURE

Traditional wearable devices have several defects, such as comfortableness for insufficient accuracy, long-term wearing etc. Thus, wellness monitoring by traditional wearable devices is not easy to be sustainable. To incur healthcare big data by comfortable health monitoring, they designed “Smart Clothing”, render less difficult unimpeded aggregation of different functional anatomy indicators of human body.

To procure pervasive computing for smart clothing method, mobile healthcare cloud chorine is fabricated by using of mobile internet, big data analytics and cloud computing. This paper introduce adjusted to facts implementation methods of smart clothing system, design details, and key technologies. Typical applications formidable by big data cloudlet and smart clothing are conferred, such as emotion care, real-time tactile interaction, and medical emergency response.

Especially, electrocardiograph signals collect information by smart clothing which is used for temper, humors, disposition mood monitoring and emotion detection or frame of mind. Finally, we highlight some of the open issues and design challenges that still need to be addressed to make smart clothing ubiquitous for wide range of applications [1].

It represents a new deeply classical architecture Bi CNN MI for rephrase or paraphrase identification (PI). Based on the penetration that PI expect compares two sentences on different levels of granularity, we acquire multi model sentence presentations using CNN and model fundamental interaction features at every level. These attractors are then input to a logistic classifier for PI.

All parametric quantity of the model (for classification, convolution and embedding) is instantly optimized for PI. To accost the shortage of preparing data, we pre-train the network in way of a novel using language modeling task. [2].

In this paper, the review of background and state-of big data. First introduce the general review related technologies and background of big data, such as Internet of Things, cloud computing, Hadoop, and data centers. Then we concentrate on 4 phases of value chain of big data, i.e. storage of data, generation of data, acquisition of data, and data analysis.

For each phase, we discussed technical challenges, the general background, and revise the modish advances and discuss the technical challenges. finally examine the various representative applications of collective intelligence, big data, including enterprise management, Internet of Things, different medial applications, online social networks, , and smart grid [3].

We present a Wearable 2.0 healthcare system to better QoS and QoE of the next generation health care system. In proposed method, washable intelligent (smart) clothing, which includes electrodes, sensors and cables, is the decisive component to integrate user's physiological data and accept the analytic thinking of results of user's health and emotionality status provided by cloud-based machine learning [4].

The US healthcare system is quickly accepting electronic health records, which will drastically increase clinical data quantity that are easily available in the form of electronic. Simultaneously, speedy progress has been made for analyzing huge quantities of data and reaping new insights from analysis-which is part of known as big data. As result shows, there are unusual chances to use big data to come down the costs of health care in the US.

We presented six use cases-that is, opportunities to reduce costs by using big data, key examples-where some of clearest high cost patients, re-admittance, triage, compensation (in case of patient's condition is worsens), optimization of treatment for diseases which affecting multiple organs systems, and adverse events. Here we discourse the various types of insights that are likely to be emerge from clinical data analytics, and the infrastructure-analytics, the various types of data needed to obtain such insights, registries, algorithms, monitoring devices, assessment scores, and so forth-that

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

Organizations will need to perform necessary analyses and to implement changes that will improve care while reducing costs. In our findings have policy implications for regulatory oversight, ways the support of research on analytics address privacy concerns and support of research on analytics [5].

The advances in IT have spectator witnessed great progression on health care technologies in several domains nowadays. However, new technologies also made health care data not only much huge but also much more complicated to process and handle. Furthermore, because the data are generated from a variety of devices in a very short span of time, the feature of these data are created quickly and stored in different formats, which can, to a large expansion, regarded as a big data problem.

To cater a more ready to hand service and environment of healthcare, In this paper purport a cyber-physical system for patient-centered health care applications and services, called HealthCPS, built on cloud and big data technologies. This system consisting a data integration layer with an incorporate standard, a data management layer for cosmopolitan storage and a data-oriented service layer and parallel computi

The results of study shows that the technology of cloud and big data analytics can be used to better the performance of the healthcare system so then humans can use various smart healthcare applications and services [6].

A small number of run risk factor come out consistently, contempt the dissimilarity of settings namely agitated confusion, instability, urinary incontinence and prescription of felon drug or 'culprit' drugs (most importantly hypnotics). Simple risk appraisal tools made of likewise variables have been presented to forecast falls with specificity and sensitivity in excess of 70%.

Although verification in a diversity of settings and in regular clinical use is not sufficient. In effect falls intervention in this population may need to use of better validated risk assessment tools, or tending to usual reversible falls risk factors in all patients [7].

Location based services, especially for vehicular location of function, are an essential integrant of most engineering technologies and applications kinder to the vehicular networks. Howsoever, because of complexness of an impulsive driving surrounding, mobility of the vehicle's movement and the effort to develop an efficient localization solution facing certain difficulties.

In this paper a convergence and hierarchical social clustering model (OHSC) is first configured to classify the vehicles into dissimilar social clusters by discovering the social connection between them. By using the output of the OHSC model, we have proposed a societal-based location algorithm (SBL) that using prediction of location to attend in global location in the vehicular networks.

The experimentation results validate the function of the OHSC model and also shows the presented SBL algorithms establishes superior localization functioning compared with the existed technique [8].

One of the ultimate goals of life science is to improve our understanding of the processes Related to disease. General health Examination is an integral part of healthcare In many countries. Identifying the patients at Risk is important for early warning and preventive intervention. The fundamental Challenge of learning a classification model for risk prediction lies in the collected dataset.

This helps to save many lives by predicting the disease at an early stage and Thereby reducing the cost of medical expenses. [9]

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?" .[10]

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

Our healthcare sector daily collects a huge data including clinical examination, vital parameters, investigation reports, treatment follow-up and drug decisions etc. But very unfortunately it is not analyzed and mined in an appropriate way. The Health care industry collects the huge amounts of health care data which unfortunately are not “mined” to discover hidden information for effective decision making for health care practitioners. Data mining refers to using a variety of techniques to identify suggest of information or decision making knowledge in database and extracting these in a way that they can put to use in areas such as decision support, Clustering, Classification and Prediction. This paper has developed a Computer-Based Clinical Decision Support System for Prediction of Heart Diseases (CCDSS) using Naïve Bayes data mining algorithm. CCDSS can answer complex “what if” queries which traditional decision support systems cannot. Using medical profiles such as age, sex, spO₂, chest pain type, heart rate, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. CCDSS is Web based, user-friendly, scalable, reliable and expandable. It is implemented on the PHP platform [11]

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not “mined” to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. This paper describes about a prototype using data mining techniques, namely Naïve Bayes and WAC (weighted associative classifier). This system can answer complex “what if” queries which traditional decision support systems cannot. Using medical profile 0073 such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established. It can serve a training tool to train nurses and medical students to diagnose patients with heart disease. It is a web based user friendly system and can be used in hospitals if they have a data ware house for their hospital. Presently we are analyzing the performances of the two classification data mining techniques by using various performance measures [12]

The data mining can be referred as discovery of relationships in large databases automatically and in some cases it is used for predicting relationships based on the results discovered. Data mining plays an important role in various applications such as business organizations, e-commerce, health care industry, scientific and engineering. In the health care industry, the data mining is mainly used for Disease Prediction. The objective our works to predict the diagnosis of heart disease with reduced number of attributes. Here fourteen attributes involved in predicting heart disease. But fourteen attributes are reduced to six attributes by using Genetic algorithm. Subsequently three classifiers like Naive Bayes, Classification by Clustering and Decision Tree are used to predict the diagnosis of heart disease after the reduction of number of attributes [13].

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 deferent word locations according to their intent importance (hence the name Deep Intent). 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. [14]

III. SYSTEM ARCHITECTURE / SYSTEM OVERVIEW

In Proposed Work, a Machine Learning i.e. Decision Tree and new convolutional neural network based unimodal disease risk prediction (CNN-UDRP) for structured data and multimodal disease risk prediction (CNN-MDRP) algorithm using unstructured and structured data from hospitals.

To the extent we know, that none of the preexistent work has focused on both types of data in the area of medical checkup big data analysis. With the development of big data analytic’s technology, more attention was paid to the prediction of diseases from the point of view of big data analyst

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

IV. SYSTEM ANALYSIS

Various studies were performed by automatically selecting the characteristics of a large amount of data to improve the exactness of the risk classification. Compared to the functions previously selected. However, these existing works have mainly considered structured data. For unstructured data, take for example, the use of the convolutional neural network (CNN) to automatically extract the characteristics of the text has already attracted great attention and has also achieved this result.

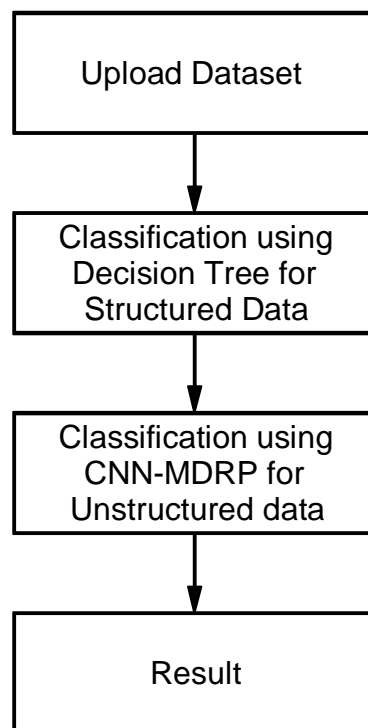


Fig1. Proposed System Architecture

IV.ALGORITHM

Decision Tree Algorithm:

Decision Tree Algorithm 1.Place impute of the dataset at the base root of the tree. 2.Split the data set into subset should be made as such a way that every subset contains data with same value for an impute. 3. Repeat step 1 and 2 on every subset until you does find its leaf or end nodes in all the branches of these tree.

CCNN-MDRP Algorithm

1. Select the specific training parameters.
2. We are use stochastic gradient method to train parameters, and finally reach the risk assessment of whether the patient suffers from disease.

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

V. COMPARATIVE RESULTS

Feasibility has following dimensions and here is a brief description in context to present System.

Financial Feasibility: The financial investment is less for creating this system. Almost all the software to be used is available on the internet. So there is no problem regarding the financial feasibility of the project. **Technical Feasibility:** It deals with the study of function, performance, and constraints like resources availability, technology, development risk that may affect the ability to achieve an acceptable system. It identifies whether the work can be done, whether technology used is compatible or not with current system. Therefore it is platform independent. The technical issues investigated during the study are as under:-

The technology for implementation of the project system is readily available. The system is capable of expansion. The presented system will provide adequate Accuracy, reliability and data security.

Behavioral Feasibility: The work of system is secured, time and storage efficient as well as accurate in providing the results and hence any user will be willing to use it.

Operational Feasibility: It is a measure of how well system solves the problems and takes the advantage of opportunities identified during scope definition and how it satisfies the requirements analysis phase of the system development.

Sr.No	Existing Method Result	Proposed Method Result
1	60 %	85 to 90 %

Table 1.comparative result

As shown in below graph proposed system has accuracy rate is extremely high than the existing system. We check with the decision tree as well as with other algorithm also and we found that system will give the best out come with CNN – MDRP so we are using this algorithm. Its gives up to the 94% good result

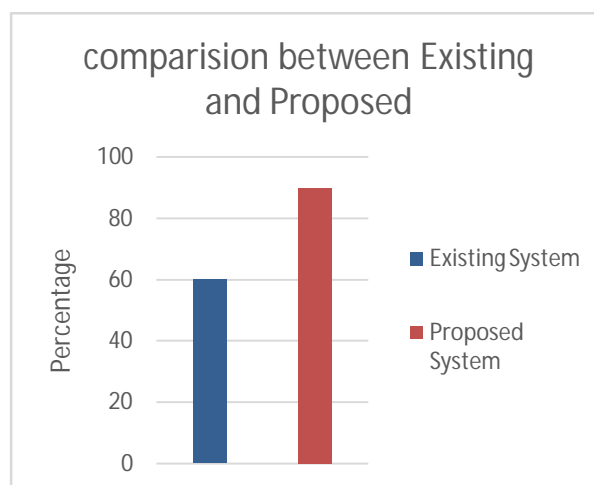


Fig.2 Graph

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

VI. CONCLUSION

In this paper, we propose a machine learning algorithm and a new multimodal disease risk prevision algorithm based on the convolutional neuronal network (CNN-MDRP) using unstructured and structured hospital data. To our knowledge, none of the existing work has focused on both types of data in the field of large medical data analysis.

Compared to many typical prediction algorithms, the predictive precision of our proposed algorithm reaches more than 90 with a faster convergence rate than that of the unimodal disease prediction algorithm based on CNNUDRP.

REFERENCES

- [1] 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.
- [2] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp
- [3] W. Yin and H. Schütze, "Convolutional neural network for paraphrase identification." in *HLT-NAACL*, 2015, pp. 901–911.
- [4] 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.
- [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] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "Healthcps: Healthcare cyber-physical system assisted by cloud and big data
- [7] 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
- [8] K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, "Localization based on social big data analysis in the vehicular networks," *IEEE Transactions on Industrial Informatics*, 2016.
- [9] S. Rajesh, G. Barani, V. Suganya K. Seenthamarai "DISEASE DETECTION BASED ON GRAPH AND DATA CLASSIFICATION," ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-4, ISSUE-9, 2017
- [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
- [11] D Ratnam, P HimaBindu , V.Mallik Sai, S.P.Rama Devi , P.Raghavendra Rao "Computer-Based Clinical Decision Support System for Prediction of Heart Diseases Using Naïve Bayes Algorithm," D Ratnam et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2384-2388
- [12] N. Aditya Sundar, P. Pushpa Latha, M. Rama Chandra3 " PERFORMANCE ANALYSIS OF CLASSIFICATION DATA MINING TECHNIQUES OVER HEART DISEASE DATA BASE," ISSN: 2250–Volume-2, Issue-3, 470 – 478
- [13] Shamsheer Bahadur Patel , Pramod Kumar Yadav, Dr. D. P. Shukla " Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques," *e-ISSN: 2319-2380, p-ISSN: 2319-2372. Volume 4, Issue 2 (Jul. - Aug. 2013), PP 61-64*
- [14] Shuangfei Zhai, Keng-hao Chang, Ruofei Zhang Zhongfei (Mark) Zhang "DeepIntent: Learning Attentions for Online Advertising with Recurrent Neural Networks," ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-4, ISSUE-9, 2017