A Parallel System Implementation of Classification and Disease Prediction on Machine Learning from Healthcare Communities

Article · February 2019

CITATIONS
0

READS
108

5 authors, including:

Swapnil Chaudhari
Marathwada Mitra Mandal's Institute of Technology, Lohgaon, Pune-47
19 PUBLICATIONS **21** CITATIONS

SEE PROFILE

A Parallel System Implementation of Classification and Disease Prediction on Machine Learning from Healthcare Communities

Sriraj Kale,Mukund Kulkarni ,Rohan Kshirsagar ,Hrishikesh Joshi,Prof.Swapnil Chaudhari

Abstract*: *To nurture admissible development, the smart city implicates a global vision that integrates artificial intelligence, decision making, information, and communication technology (ICT) and the internet-of-things (IoT) together. In this project, the topic of disease prediction and diagnosis in smart healthcare is analyzed. Due to data processed in biomedical and healthcare communities, accurate study of medical data benefits early disease recognition, patient care, and community services. When the quality of medical data is insufficient the exactness of analysis is reduced. Moreover, different regions demonstrate unique appearances of certain regional diseases, which may result in diminishing the prediction of disease outbreaks. In the proposed system, it provides machine learning algorithms for effective prediction of various disease prevalence in disease-frequent societies and predicts the waiting time for every treatment task for each patient as well as a Hospital Queuing Recommendation (HQR) system is developed for recommending treatment task sequence with respect to expected waiting time. It demonstrates on a regional chronic illness of cerebral infarction. Using structured and unstructured data from the hospital it uses Machine Learning algorithms like the Decision Tree algorithm and the KNN algorithm. To the foremost of our knowledge in the area of medical big data analytics, none of the existing systems and its purpose focused on both data types. Compared to various typical estimated algorithms, the calculation exactness of our proposed algorithm outreach 94.8% with a convergence speed which is more accurate, faster than that of the CNN-based uni-modal disease risk prediction (CNN-UDRP) algorithm. In addition, challenges in the deployment of disease diagnosis in healthcare have been discussed.

Index Terms: ***KNN, PTTP, Disease Prediction.***

I. INTRODUCTION

In proposed system we focus on helping patients complete their treatment tasks in a predictable time and helping hospitals schedule each treatment task queue and avoid overcrowded and ine ective queues. We use massive realistic data from various hospitals to develop a patient treatment time consumption model. The realistic patient data are analyzed carefully and rigorously based on important parameters, such as patient treat-ment start time, end time, patient age, and detail treatment content for each di erent task. We identify and calculate di erent waiting times for di erent patients based on their conditions and operations performed during treatment.

II. MOTIVATION

The motivation behind this report is to give a gritty review of the product item "Enhancing Disease Prediction by Machine Learning Approaches". This record portrays the task's intended interest group and its UI, equipment and programming prerequisites. It characterizes how the customer, group and gathering of people see the item and its usefulness.

The healthcare problem of chronic diseases has been increased in many countries. Therefore, it is essential to perform risk assessments for chronic diseases. Motivated by this problem, Support Vector Machine algorithm Naive Bayesian algorithm are been proposed for disease prediction using unstructured and structured data, respectively.

In proposed system we focus on helping patients complete their treatment tasks in a predictable time and helping hospitals schedule each treatment task queue and avoid overcrowded and ineffective queues. We use massive realistic data from various hospitals to develop a patient treatment time consumption model. The realistic patient data are analyzed carefully and rigorously based on important parameters, such as patient treat-ment start time, end time, patient age, and detail treatment content for each different task. We identify and calculate different waiting times for different patients based on their conditions and operations performed during treatment.

.

III. OBJECTIVES

- To reduce waiting time of patents in hospital.
- To overcome the problem of for carrying documents, reports, etc. i.e. paper less work.
- To reduce manpower in every stage.
- To nd easily nearest doctors or hospitals.
- To predict waiting time of each treatment task.
- A Proposed system to analyze Patient Treatment Time Prediction (PTTP) algorithm to predict the waiting time for each treatment task for a patient.

- HQR calculates and predicts and recommends plan for the patient.

IV. SOFTWARE SPECIFICATION REQUIREMENTS

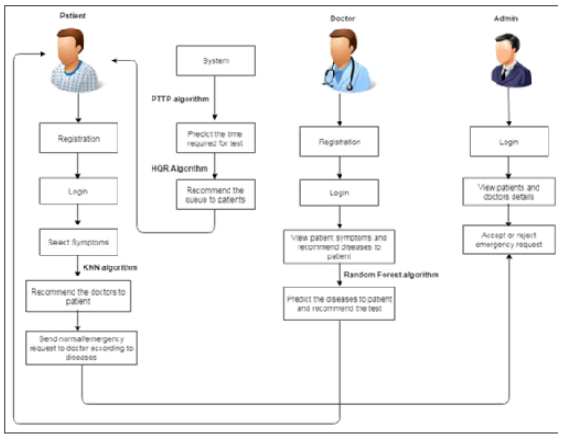
A software requirements specification (SRS) is a document that is created when a detailed description of all aspects of the software to be built must be specified before the project is to commence. It is important to note that a formal SRS is not always written. In fact, there are many instances in which e ort expended on a SRS might be better spent in other software engineering activities. to the data sharing model applying the SBIBD, multiple participants can form a group to efficiently share the outsourced data. Subsequently, each group member performs the key agreement to derive a common conference key to ensure the security of the outsourced group data. Note that the common conference key is only produced by group members. We propose a new Convolutional Neural Network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital. The prescribed tests waiting time of each treatment task is obtained by the PTTP model, which is the sum of all patients' probable treatment times in the current queue. An HQR system is proposed based on the predicted waiting time. A treatment recommendation with an e cient and convenient treatment plan and the least waiting time is recommended for each patient. The PTTP algorithm and HQR system are extensive hospital data are stored in the database

V. IMPLEMENTATION

A. Figures and Tables

1. System Design:

We propose a new convolution neural network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital. On the basis of decision tree algorithm we predict the diseses that patient caused.The prescribed tests waiting time of each treatment task is obtained by the PTTP model, which is the sum of all patients' probable treatment times in the current queue. An HQR system is proposed based on the predicted waiting time. A treatment recommendation with an e cient and convenient treatment plan and the least waiting time is recommended for each patient. The Patient treatment time prediction algorithm and Hospital queue recommendation system are extensive hospital data are stored in the database.



2. **Dataset And Model Description:**

In this section, we describe the hospital datasets we use in this study. Furthermore, we provide disease risk prediction model and evaluation methods.

A:Hospital Data :

The hospital dataset used in this study contains real-lifehospital data, and the data are stored in the data center. To protect the patient's privacy and security, we created a security access mechanism. The data provided by the hospital include EHR, medical image data and gene data. We use a three year data set from 2013 to 2015. Our data focus on inpatient department data which included 31919 hospitalized patients with 20320848 records in total. The inpatient department data is mainly composed of structured and unstructured text data. The structured data includes laboratory data and the patient's basic information such as the patient's age, gender and life habits, etc. While the unstructured text data includes the patient's narration of his/her illness, the doctor's interrogation records and diagnosis, etc. As shown in Table I, the real-life hospital data collected from central China are classified into two categories, i.e., structured data and unstructured text data In order to give out the main disease which affect this region, we have made a statistics on the number of patients,the sex ratio of patients and the major disease in this region every year from the structured and unstructured text data, the statistical results are as shown in Table II. From Table II, we can obtain that the proportion of male and female patients hospitalized each year have little difference and more patients admitted to the hospital in 2014.

TABLE I

ITEM TAXONOMY IN CHINA HOSPITAL DATA

Data category	Item	Description
Structured data	Demographics of the patient	Patient's gender, age, height, weight, etc.
	Living habits	Whether the patient smokes, has a genetic history, etc.
	Examination items and results	Includes 682 items, such as blood, etc.
	Diseases	Patient's disease, such as cerebral infarction, etc.
Unstructured text data	Patient's readme illness	Patient's readme illness and medical history
	Doctor's records	Doctor's interrogation records

TABLE II
INITIAL STATISTICS FROM HOSPITAL DATA IN WUHAN,CHINA

Statistics	2013	2014	2015
Number of inpatients	7265	24756	10552
Males	42.88%	50.36%	57.60%
Females	57.12%	49.64%	42.40%
Proportion of patients with cerebral infarction	1.47%	1.01%	1.66%
Proportion of hypertensive patients	1.06%	1.04%	1.98%
Proportion of diabetics	1.17%	0.99%	1.99%

B. Disease Risk Prediction:

From Table II, we obtain the main chronic disease in this region. The goal of this study is to predict whether a patient is amongst the cerebral infarction high-risk population according to their medical history. More formally, we regard the risk prediction model for cerebral infarction as the supervised learning methods of machine learning, i.e., the input value is the attribute value of the patient, $X = (x_1; x_2; \ ; x_n)$ which includes the patient's personal information such as age, gender, the prevalence of symptoms, and living habits (smoking or not) and other structured data and unstructured data.

The output value is C, which indicates whether the pa- tient is amongst the cerebral infarction high-risk population. $C = fC_0; C_1g$, where, C_0 indicates the patient is at high- risk of cerebral infarction, C_1 indicates the patient is at low- risk of cerebral infarction. The following will introduce the dataset, experiment setting, dataset characteristics and learning algorithms briefly.

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.

In the experiment setting and dataset characteristics, we select 706 patients in total as the experiment data and randomly divided the data into training data and test data. The ratio of the training set and the test set is 6:1 [22], [23] , i.e., 606 patients as the training data set while 100 patients as the test data set. We use the C++ language to realize the machine learning and deep learning algorithms and run it in a parallel fashion by the use of data center. In this paper, for S-data, according to the discussion with doctors and Pearson's correlation analysis, we extract the patient's demographics characteristics and some of the characteristics associated with cerebral infarction and living habits (such as smoking). Then, we obtain a total of patient's 79 features. For T-data, we first extract 815073 words in the text to learn Word Embedding. Then we utilize the independent feature extraction by CNN.

We will introduce machine learning and deep learning algorithms used in this work briefly. For S-data, we use three conventional machine learning algorithms, i.e., Naive Bayesian (NB), K-nearest Neighbour (KNN), and Decision Tree (DT) algorithm [24], [25] to predict the risk of cerebral infarction disease. This is because these three machine learning methods are widely used [26]. For T-data, we propose CNN-basedunimodal disease risk prediction (CNN-UDRP) algorithm to predict the risk of cerebral infarction disease. In the remaining of the paper, let CNN-UDRP(T-data) denote the CNN-UDRPalgorithm used for T-data. For S&T data, we predict the risk of cerebral infarction disease by the use of CNN-MDRPalgorithm, which is denoted by CNN-MDRP(S&T-data) for the sake of simplicity. In the following section, the details about CNN-UDRP(T-data) and CNN-MDRP(S&T data) will be given.

C. *Evaluation Methods:*

For the performance evaluation in the experiment. First, we denote TP , FP , TN and FN as true positive (the number of instances correctly predicted as required), false positive (the number of instances incorrectly predicted as required), true negative (the number of instances correctly predicted as not required) and false negative (the number of instances incorrectly predicted as not required), respectively. Then, we can obtain four measurements: accuracy, precision, recall and F1-measure as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad ; \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-Measure} = \frac{2 \ \text{Precision} \ \text{Recall}}{\text{Precision} + \text{Recall}} \ ;$$

where the F1-Measure is the weighted harmonic mean of the precision and recall and represents the overall performance.

In addition to the aforementioned evaluation criteria, we use receiver operating characteristic (ROC) curve and the area under curve (AUC) to evaluate the pros and cons of the classifier. The ROC curve shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR), where the TPR and FPR are defined as follows:

$$TPR = \frac{TP}{TP + FN} \ ; FPR = \frac{FP}{FP + TN}$$

If the ROC curve is closer to the upper left corner of the graph, the model is better.

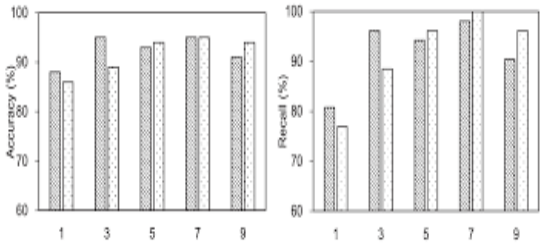
The AUC is the area under the curve. When the area is closer to 1, the model is better. In medical data, we pay more attention to the recall rather than accuracy. The higher the recall rate, the lower the probability that a patient who will have the risk of disease is predicted to have no disease risk.

VI. EXPERIMENTAL RESULTS

In this section, we discuss the performance of CNN-UDRPand CNN-MDRP algorithms from several aspects, i.e.,the run time, sliding window, iterations and text feature.

A. Run Time Comparison

We compare the running time of CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms in personal computer (2core CPU, 8.00G RAM) and data center (6core*2*7=84core CPU, 48*7=336G RAM). Here, we set the same CNN itera- tions which are 100 and extract the same 100 text features



(a) (b)

As shown in Fig. 2, for CNN-UDRP (T-data) algorithm, the running time in data center is 178.5s while the time in personal computer is 1646.4s. For CNN-MDRP (S&T-data) algorithm, its running time in data center is 178.2s while the time in personal computer is 1637.2s. That is, the running speed of the data center is 9.18 times on the personal computer. Moreover, we can see the running time of CNN-UDRP (T- data) and CNN-MDRP (S&T-data) are basically the same from the figure, i.e. although the number of CNN-MDRP (S&T- data) features increase after adding structured data, it does not make a significant change in time. The later experiments are based on the running results of the data center.

B. . *Effect of Sliding Window (Word Number)*

When taking convolution of CNN, we need to confirm the number of words for sliding window first. In this experiment, the selected number of words for the sliding window are 1, 3, 5, 7 and 9. The iterations of CNN are 200 and the size of convolution kernel is 100. As shown in Fig. 3, when the number of words for the sliding window are 7, the accuracy and recall of CNN-UDRP (T-data) algorithm are 0.95 and 0.98, respectively. And the accuracy and recall of CNN-MDRP(S&T-data) algorithm are 0.95 and 1.00. These results are all higher than we choose other number of words for sliding window. Thus, in this paper, we choose the number of words for sliding window are 7.

C. Effect of Iterations:

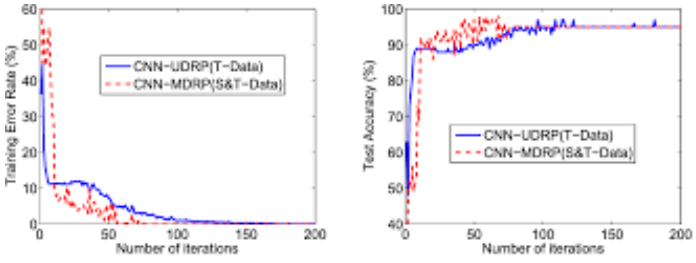
We give out the change of the training error rate and test accuracy along with the number of iterations. As shown in Fig. 4, with the increase of the number of iterations, the training error rate of the CNN-UDRP (T-data) algorithm decreases gradually, while test accuracy of this method increases. The CNN-MDRP (S&T-data) algorithm have the similar trend in terms of the training error rate and test accuracy. In Fig, we can also obtain when the number of iterations are 70, the training process of CNN-MDRP (S&T-data) algorithm is
already stable while the CNN-UDRP (T-data) algorithm is still not stable. In other words, the training time of MDRP(S&T data) algorithm is shorter, i.e. the convergence speed of CNN- MDRP (S&T-data) algorithm is faster.

VII. ANALYSIS OF OVERALL RESULTS

In this section, we describe the overall results about S-dataand S&T-data.

A. Structured Data (S-data):

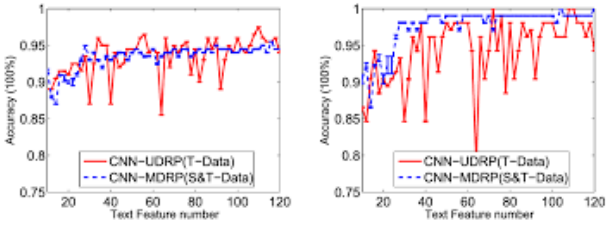
For S-data, we use traditional machine learning algorithms, i.e., NB, KNN and DT algorithm to predict the risk of cerebral infarction disease. NB classification is a simple probabilistic classifier. It requires to calculate the probability of feature attributes. In this experiment, we use conditional 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.

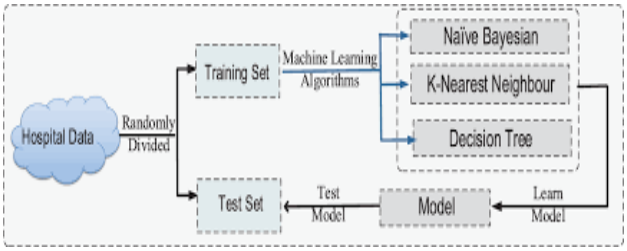


B.Structured and Text Data (S&T-data):

According to the discussion in Section IV, we give out the accuracy, precision, recall, F1-measure and ROC curve underCNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithm- s. In this experiment, the selected number of words is 7 and the text feature is 100. As for CNN-UDRP (T-data) and CNN- MDRP (S&T-data) algorithms, we both run 5 times and seek the average of their evaluation indexes. From the Fig. 8, the accuracy is 0.9420 and the recall is 0.9808 under CNN-UDRP(T-data) algorithm while the accuracy is 0.9480 and the recall is 0.99923 under CNN-MDRP (S&T-data) algorithm. Thus, we can draw the conclusion that the accuracy of CNN-UDRP(T-data) and CNN-MDRP (S&T-data) algorithms have little difference but the recall of CNN-MDRP (S&T-data) algorithm is higher and its convergence speed is faster. In summary, the performance of CNN-MDRP (S&T-data) is better than CNN- UDRP (T-data).

In conclusion, 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., hyperlipidemia, only a few features of structured data can get a good description of the disease, resulting in fairly good effect of disease risk prediction [33]. But for a complex disease, such as cerebral infarction mentioned in the paper, only using features of structured data is not a good way to describe the disease. As seen from Fig. 7(a) and Fig. 7(b), 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





CONCLUSION &FUTURE WORK

Chronic or heart disease has increased, a new conventional neural network based multi-modal disease risk prediction (CNNMDRP) algorithm in which structured and unstruc-tured data from hospital is being used. In this structured and unstructured data, the personal information and detail history of the patient is being stored. In this CNN-MDRP both data are being used for predicting the chronic disease in that particular patient. In unstructured data patients may have missing data. So, the missing data of that particular patient can also retrieve through the genetic algorithm. The featured from unstructured data are been extracted correctly. Then the extracted features are structured data. Both Structured data and extracted structured data are used for predicting the exact chronic disease with Naive Bayes classi er and the SVM classi er. In the proposed system, it pro-vides machine learning algorithms for e ective prediction of various disease occurrences in disease-frequent societies and predicts the waiting time for every treatment task for each patient as well as a Hospital Queuing Recommendation (HQR) system is developed for recommending treatment task sequence with respect to expected waiting time.

REFERENCES

[1 [1J. 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.

[1 [2]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-costpatients," *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014.

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

[1 [4]Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "Health- cps: Healthcare cyber-physical system assisted by cloud and big data," *IEEE Systems Journal*, 2015.

1 [5]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.

1 [6]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.

[1 [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.

[1 [8]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.

[9]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.

[1[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.