MEDICAL RECORDS SENSITIVE DETAILS RETRIEVAL: ANALYSIS, OPTIMIZATION AND ENCRYPTION

B. Ayeshwarya, B.E.,
PG Student, Department of Computer Science and Engineering
C.Abdul Hakeem College of Engineering and Technology, Melvisharam, Vellore, Tamil Nadu, India

Mr. C. Kotteeswaran, M.E., Ph.D.,
Associate Professor, Department of Computer Science and Engineering
C.Abdul Hakeem College of Engineering and Technology, Melvisharam, Vellore, Tamil Nadu, India

Abstract

There is a sustainable growth in Big Data in biomedical and healthcare communities, which accurately analyzes the benefits of medical data such as early disease detection, patient care and community services. However, when the quality of medical data is insufficient, the accuracy of the analysis will be reduced. In addition, different regions exhibit unique characteristics of certain regional diseases, which may undermine the predictions of disease outbreaks. In this paper, we are simplifying machine learning algorithms to effectively predict the chronic heart disease in disease-prone communities. We are considering a predictive model of real hospital data collected from UCI repository. To overcome data inefficiencies, we used a latent factor model to reconstruct the lost data. We proposed a new convolutional neural network (CNN) algorithm that uses structured and unstructured data from hospitals. As far as we know, the existing work does not focus on two types of data in the field of medical big data analysis. Compared with several typical prediction algorithms, the proposed method has a prediction accuracy of 83% and a faster convergence rate than the existing algorithms. For security reasons, we used the DES algorithm to encrypt the entire data set.

Keywords: *Classification, Machine Learning, Data Analysis, Risk Prediction, Encryption.*

Introduction

With the advancement of enormous information investigation innovation, more consideration has been paid to disease prediction from the perspective of data analytics. Among several life-threatening ailments, heart disease has a great impact over recent times in medical research. The diagnosis of heart disease at early stage of patient with the goal that further treatment can be made successful involving low risk. The analysis of heart disease is normally founded on signs, indications and physical examination of the patient. The major causes of heart disease includes smoking habit, body cholesterol level, family ancestry of heart disease, weight, high blood pressure. Heart disease have several symptoms like chest pain or discomfort, upper body pain or discomfort in the arms, back, neck, jaw, or upper stomach, shortness of breath, nausea, lightheadedness, or cold sweats.

Several health conditions, your lifestyle, and your age and family history can increase your risk for heart disease. These are called risk factors. About half of all Americans (47%) have at least one of the three key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Some of the risk factors for heart disease cannot be controlled, such as your age or family history. But you can take steps to lower your risk by changing the factors you can control. Some of the Heart Disease Facts are as follows: 1) Heart disease is the leading cause of death for both men and women. About 630,000 Americans die from heart disease each year - that’s 1 in every 4 deaths. 2) In the

United States, someone has a heart attack every 40 seconds. Each minute, someone in the United States dies from a heart disease-related event. 3) Coronary heart disease is the most common type of heart disease, killing more than 370,000 people annually. 4) Coronary heart disease alone costs the United States $108.9 billion each year. This total includes the cost of health care services, medications, and lost productivity. 5) Every year about 735,000 Americans have a heart attack. Of these, 525,000 are a first heart attack and 210,000 happen in people who have already had a heart attack.

Heart problems can be avoided in the future by adopting a healthy lifestyle today. Here are some simple ways to prevent heart disease. 1. Don't smoke or use tobacco, 2. Eat a heart-healthy diet, 3. Exercise for about 30 minutes on most days of the week, 4. Get enough quality sleep, 5. Maintain a healthy weight, 6. Manage stress. These superior ways can help improve your health.

Literature Survey

At present, although the significance of big data has been generally recognized, many individuals still have various opinions on its definition. The following definitions may help us have a better understanding on the profound social, economic, and technological implications of big data. In 2010, Apache Hadoop defined big data as “datasets which could not be captured, managed, and processed by general computers within a worthy scope.” By the definition, in May 2011, McKinsey and Company, a worldwide consulting agency announced Big Data as the next frontier for development, rivalry, and profitability. In addition, NIST defines big data as “Big data shall mean the data of which the data volume, acquisition speed, or data representation limits the capacity of using conventional relational methods to conduct adequate analysis or the data which may be effectively processed with critical horizontal zoom technologies” [1]. Jensen [2], first introduced the typical content of a generic EHR system and focused on how data driven knowledge discovery on cohort-wide health data can fill and assist informed clinical decision making. Also described how the integration of EHR and genetic data, together with systems biology approaches, can facilitate genotype–phenotype associated researches. Finally discussed some of the structural and political challenges that are facing EHR adoption.

The network selection is an essential step to the realization of multimode communications in heterogeneous vehicular telematics that utilize multiple access technologies and multiple radios in a collaborative manner. A well enhanced network for the fundamental technological requirement of multimode communications in heterogeneous vehicular telematics was identified. A dynamic network selection method which satisfies the QoS requirements of different terminal’s applications and also ensures the efficient utilization and fair allocation of heterogeneous network resources in a global sense was developed [3]. The proposed network selection strategy provides better global performance when compared with the utility function approach with greedy optimization.

A Wearable 2.0 healthcare system to improve QoE and QoS of the next generation healthcare system was proposed by Chen et.al [4]. The washable smart clothing consists of sensors, electrodes, and wires, which is the essential component to collect users’ physiological data and receive the analysis results of users’ health and emotional status provided by cloud-based machine intelligence. Typical applications powered by smart clothing and big data clouds are introduced, such as medical emergency response, emotion care, illness diagnosis, and real-time tactile interaction. Especially, electrocardiograph signals collected by smart clothing are used for mood monitoring and emotion detection [5]. Based on the four tasks: collection, communication, analysis and feedback of emotion the architecture of emotion communication system was created [6]. At the same time, to maintain the synchronization requirements of the communication for both the two sides when the emotion is transmitted as a kind of multimedia information, an emotion communication protocol was proposed, which provides a high-level reliable support for the acknowledgment of emotion communications. Finally, the real-time performance of a speech emotion communication system based on the pillow robot is analyzed, and the feasibility and effectiveness of realizing the emotion communications are highlighted. Qiu et.al [7] proposed a heterogeneous solution such that the total cost is minimized while the timing constraint is satisfied with a guaranteed confidence probability. The real-time embedded systems are involved for high-level synthesis of the functional units. The solutions can be implemented for both hard real-time and soft real-time systems. Optimal solutions may be realized as tree or simple path. The solutions for the algorithm of heterogenous system attains the confidence probability with an average reduction on total cost satisfying timing constraints. Patients information are recorded in the EHR for the reduction of cost in medical studies.

Qiu et.al [8] proposed a probability-based bandwidth model in a telehealth cloud framework, which helps cloud broker to provide a high performance allocation of computing nodes and links. This brokering system 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 use of transmission in a telehealth context. By analyzing the features of data processing with medical applications, a decentralized data coherence protocol to solve the performance issues by current design was developed. Their model measures the bandwidth consumption between any node pair in cloud so that the bandwidth can be calculated in each interval. 28 papers on risk factors were identified, with 15 excluded from further analysis [9]. The nature of risk factors in hospital inpatients and the ability to identify high-risk patients was the major process in the design of future falls prevention interventions. They may also be appropriate to other facilities, which provide care for post acute patients, such as Intermediate Care units in the UK or talented nursing facilities in the US. The patient risk prediction problem in the context of active learning with relative similarities is investigated [10]. Active learning has been applied to solve real problems which is to query absolute questions. In a medical application where the objective is to predict the risk of patients on certain disease using Electronic Health Records (EHR).

Several researches have been conducted by selecting the characteristics automatically from a large number of data to improve the accuracy of risk classification, comparing with the existing selected characteristics. The existing works focused mostly on structured data. For unstructured data using convolutional neural network (CNN) to extract text characteristics has created much impact and also achieved greater results. The main advantage of our proposed method is encrypting the whole data set using DES algorithm. Through the experiment, we draw a conclusion that the performance of CNN is better than other existing methods after encryption.

The rest of this paper is organized as follows: Section III describes the various methods used in this paper. Section IV provides the experimental discussions about the performance of the proposed system. Section V concludes the results of this paper.

Methods

Implementation

The Dataset is uploaded into the database by the admin. By giving the user id, risk level is shown. Firstly data is classified into two parts, one is structured data and another is unstructured data. For both the things, risk level is calculated based on the constraints. Then the overall risk level is calculated by comparing the structured and unstructured data. For both the low level and high level, it recommends the prescription. If any fields or data is missing, we can update the details, then it calculate the risk level.

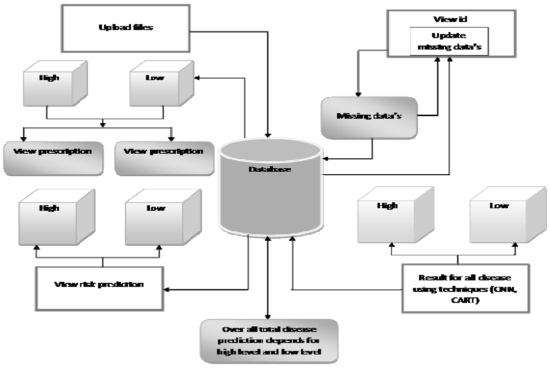


Figure 1: Structure of Risk Prediction

Data Repository

Admin needs to login with username and password. If both match, then he/she will be considered as valid person. After login, Admin has to upload disease datasets which has 13 attributes of the patients. The uploaded dataset can be viewed by the admin and he can edit the missing fields in the dataset. Admin has the capability to maintain all user details.

Data Classification

Data can be classified into Structured data and unstructured data. The Structured data is based upon the Laboratory report. And the Unstructured data is retrieved from the dataset.

Classification of Structured Data

The structured data includes laboratory data and the patient's basic information such as the patient's age, gender and life habits, etc. Structured data (S-data): use the patient's structured data to predict whether the patient is at high-risk of heart disease or Low-risk of heart disease. The features number of structured text data extracted by using Decision tree algorithm. Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, decision tree algorithm can be used for solving regression and classification problems too. The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data (training data).

Classification of Unstructured Data

While the unstructured text data (chol, fbs, restecg, thalach, exang) includes the patient's narration of his/her illness, the doctor's interrogation records and diagnosis, etc. Unstructured data (U-data): use the patient's unstructured text data to predict whether the patient is at high-risk or Low-risk of Heart disease. The features number of unstructured text data are extracted by using CNN algorithm. The Risk level of the unstructured data can be predicted using CNN algorithm.

Data Analysis

We are comparing structured and unstructured data, and then the risk analysis is predicted. Both the High and Low risk level has been predicted. CNN and decision tree algorithm used here to predict the Low level and high level risk for both the structured and unstructured data.

Risk Result

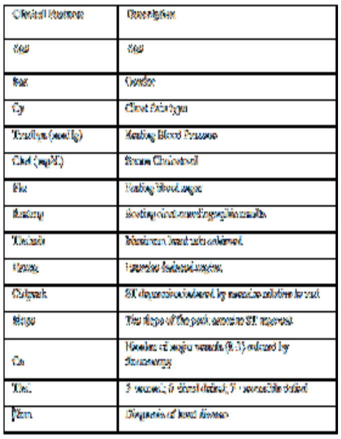
Finally the result has been shown to the user at which level they are. The prescription for both the high level and low level of heart disease has been shown to the user.

Overall Risk Result

Graphs has been generated based upon the number of patients, high risk and low risk. Efficiency is calculated depends on the true positive (TP), true negative (TN), false positive (FP) and false negative (FN). So that the accuracy is calculated by using the following formula,

Accuracy= (TP+TN)/TP+FP+TN+FN

Table 1 Attributes used for the experiment



Clinical Features	Description
Age	Age
Sex	Gender
Cp	Chest Pain type
Trestbps (mm Hg)	Resting Blood Pressure
Chol (mg/dl)	Serum Cholesterol
Fbs	Fasting blood sugar
Restecg	Resting electrocardiographic results
Thalach	Maximum heart rate achieved
Exang	Exercise induced angina
Oldpeak	ST depression induced by exercise relative to rest
Slope	The slope of the peak exercise ST segment
Ca	Number of major vessels (0-3) colored by fluoroscopy
Thal	3 = normal; 6 = fixed defect; 7 = reversible defect
Num	Diagnosis of heart disease

Performance Evaluation Criteria

Four different metrics i.e. accuracy, precision, recall, and F1-measures are used to evaluate the performance of the proposed system. 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 metrics: accuracy, precision, recall, and F1- measure as follows:

Accuracy = TP+TN/TP+TN+FP+FN

Precision = TP/TP+FP

Recall = TP/TP+FN

FI-Measure = 2*Precision*Recall/ Precision + Recall

Accuracy, Precision, Recall, and F1-Measure are used to express the success of predicting the risk of heart disease.

Convolutional neural network

In neural networks, Convolutional neural network (ConvNets or CNNs) is one of the main categories to do visual imagery. Objects detections, recognition faces etc., are some of the areas where CNNs are widely used. CNNs use a variation of multilayer perceptrons designed to require minimal preprocessing. They are also known as shift invariant or space invariant artificial neural networks (SIANN). CNNs use relatively little pre-processing compared to other image classification algorithms. A convolutional neural network consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of convolutional layers, activation function, pooling layers, fully connected layers and normalization layers.

Convolutional Layers

Convolutional layers apply a convolution operation to the input, passing the result to the next layer. The convolution imitates the reaction of an individual neuron to visual improvements.

Activation Function

In a neural network, the activation function is responsible for transforming the summed weighted input from the node into the activation of the node or output for that input.

Pooling Layers

Convolutional networks may include local or global pooling layers, which combine the outputs of neuron clusters at one layer into a single neuron in the next layer. For example, max pooling uses the maximum value from each of a cluster of neurons at the prior layer. Another example is average pooling, which uses the average value from each of a cluster of neurons at the prior layer.

Fully Connected Layers

Fully connected layers connect every neuron in one layer to every neuron in another layer. It is in principle the same as the traditional multi-layer perceptron neural network (MLP). The flattened matrix goes through a fully connected layer to classify the risk.

Normalization Layers

Many types of normalization layers have been proposed for use in ConvNet architectures, sometimes with the intentions of implementing inhibition schemes observed in the biological brain.

Decision Trees

The Decision tree is a classification strategies in which classification is done for dividing the patients according to their risk classification prediction. The decision tree is a schema like a tree structure that gatherings instances by arranging them in context of the feature values. Each and every node in a decision tree depicts the features to be classified according to the attributes mentioned in the Table 1. Decision tree makes the standard for the classification of the data set. We use CART algorithm for classifying the risk level of the patients. CART stands for Classification and Regression Trees. It was invented by Breiman in 1984. The classification tree development by CART is depends on binary separating of the properties. CART also based on Hunt’s algorithm and can be executed serially. Gini index is utilized as splitting measure in picking the splitting attribute. CART is not quite same as other Hunt’s based algorithm.

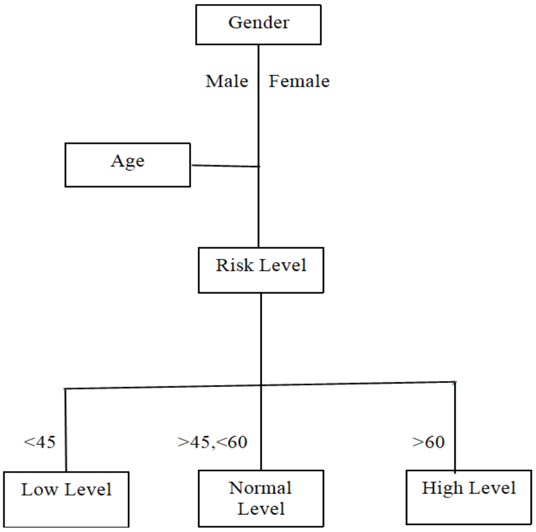


Figure 2: Decision Tree Example

Des Encryption

The Data Encryption Standard (DES) is a symmetric-key method of data encryption published by the National Institute of Standards and Technology (NIST). DES works by using the same key to encrypt and decrypt a message, so both the sender and the receiver must know and use the same private key. DES was adopted by the U.S in the early 1970s by researchers at IBM. DES was quickly adopted by industries such as financial services, where the need for strong encryption is high.

Result and Discussions

With the focus of predicting the risk level of the heart disease for the patient records, we define an effective machine learning algorithm. The dataset consists of 215 instances and each instance contains 15 attribute features. The structured risk includes ID, age, gender and name. The unstructured risk includes cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, num. The overall risk level is predicted for the patient using the details of the Structured risk and Unstructured risk.

Figure 3, represents the recall performance of the existing system and the proposed method. The proposed approach demonstrates the improvement in the performance than existing methods. Recall is the ratio of correctly predicted positive observations to the all observations in genuine class - yes.

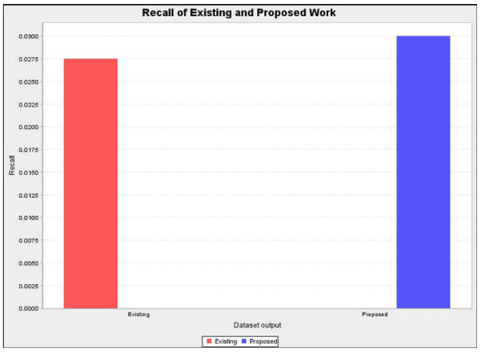


Figure 3: Performance Improvement of Recall

Figure 4, represents the precision performance of the existing system and the proposed method. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The proposed method well suits to process in the encrypted format.

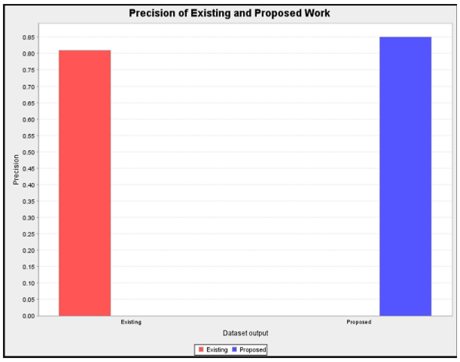


Figure 4: Performance Improvement of Precision

Figure 5, represents the F1 Measure performance of the existing system and the proposed method. F1 Measure is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

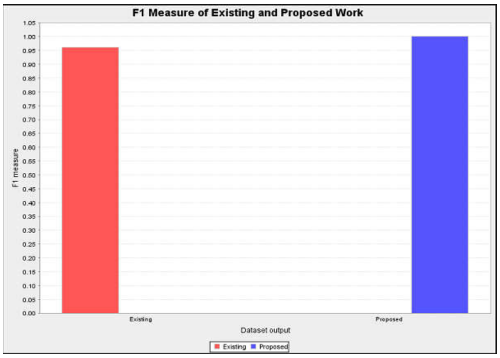


Figure 5: Performance Improvement of F1 Measure

Figure 6, represents the accuracy performance of the existing system and the proposed method. Accuracy is the most important performance measure and it is simply a ratio of correctly predicted observation to the total observations. The proposed method visualize a higher accuracy for the encrypted approach.

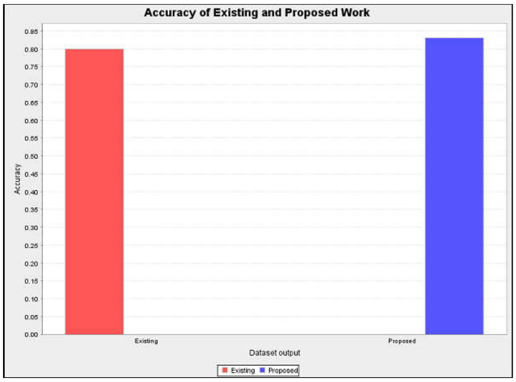


Figure 6: Performance Improvement of Accuracy

Conclusion

In this paper, we proposed a new convolutional neural network algorithm using unstructured and Decision tree for structured data from hospital data. The structured data classification is based on Classification and Regression Tree (CART) approach. The unstructured data classification is based on Convolutional Neural Network (CNN) approach. The process consists of three main steps. The first step is to classify the structured data and unstructured data. The second step is to predict and analysis the risk level of the patient. The third step is to generate the overall risk result. Compared with several typical prediction algorithms, the proposed method has a prediction accuracy of 83% and a faster convergence rate than the existing algorithms. For security reasons, we used the DES algorithm to encrypt the entire data set which avoids the risk of privacy issues for the patient's record. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

References

1. M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” Mobile Networks and Applications, vol. 19, no. 2, pp. 171–209, 2014.
2. 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.
3. 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.
4. 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.
5. 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.
6. M. Chen, P. Zhou, G. Fortino, “Emotion Communication System,” IEEE Access, DOI: 10.1109/ACCESS.2016.2641480, 2016.
7. 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.
8. J. Wang, M. Qiu, and B. Guo, “Enabling real-time information service on tele health system over cloud-based big data platform,” Journal of Systems Architecture, vol. 72, pp. 69–79, 2017.
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 and Knowledge Discovery, vol. 29, no. 4, pp. 1070–1093, 2015.