HEALTH RISK PREDICTION BY MACHINE LEARNING OVER DATA ANALYTICS

PRABHU. T[1], DARSHANA. J[2], DHARANI KUMAR. M[3], HANSAA NAZREEN. M[4]

[1]Assistant Professor Department of ECE, Department of Electronics and Communication Engineering, SNS College of Technology, Coimbatore, Tamil Nadu, India
[2, 3 4]Final Year ECE Students, Department of Electronics and Communication Engineering, SNS College of Technology, Coimbatore, Tamil Nadu, India

---***---

Abstract - *The monumental value of health care, particularly for chronic disorder treatment, is quickly becoming unmanageable. This crisis has intended the drive towards preventative remedy, where the primary concern is recognizing disease risk and taking action at the earliest signs. However, universal testing is neither time nor value economical. We propose, a Collaborative Assessment and Recommendation, which relies on a person's medical history, lifestyle habits and Big Data of similar records in order to predict future diseases risks. This combines collaborative filtering methods with clustering to predict each patient's greatest 3disease risks based on their own medical history and that of similar patients. We also describe an Iterative version, which incorporates ensemble concepts for improved performance. We thereby propose a new Convolutional Neural Network Based Multimodal Disease Risk Prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital. To the most effective of our data, 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 would reach high percentiles with a convergence speed which is faster than that of the CNN-Based Unimodal Disease Risk Prediction (CNN-UDRP) algorithm. These novel systems need no specialised data and supply predictions for medical conditions of all types in a very single run. We present experimental results on a large Medicare dataset, demonstrating that this performs well at capturing future disease risks using machine learning algorithms.*

Key Words: Data Analytics, Structured and Unstructured Data, Convolution Neural Network, Prediction Accuracy, Convergence Speed.

1. INTRODUCTION

According to a report, 50% of Americans have one or more chronic diseases, and 80% of American medical care fee is spent on chronic disease treatment. With the development of living standards, the incidence of chronic disease is increasing. The United States has spent a mean of 2.7 trillion USD annually on chronic malady treatment. This quantity includes eighteen of the whole annual Gross Domestic product (GDP) of the United States(US). The setback of chronic diseases is additionally important in several other countries. Therefore, it's essential to perform risk assessments for chronic diseases.

With the expansion in medical information, collecting electronic health records (EHR) is increasingly convenient. Besides, a bio-inspired high-performance heterogeneous vehicular telematics 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. One of the applications is to spot risky patients which may be used to cut back medical value since risky patients usually need dearly-won attention. Prediction victimization ancient malady risk models typically involves a machine learning algorithm program (e.g., logistic regression and regression analysis, etc.), and especially a supervised learning algorithm by the use of training knowledge with labels to instruct the model. In the examine set, patients can be classified into groups of either high-risk or low-risk. These models area unit valuable in clinical things and area unit widely studied. However, these schemes have the following characteristics and defects. The data set is often little, for patients and diseases with specific conditions; the characteristics area unit chosen through expertise. However, these pre-selected characteristics maybe not satisfy the changes in the disease and its in quencing factors.

2. EXISTING SYSTEM

As health care applications generate great deal of knowledge that varies with reference to its volume, variety, velocity, veracity, and value, there's is an impending demand of economical mining techniques for context-aware retrieval and process of this category of knowledge. So, it proposes a brand new fuzzy rule-based classifier to produce Healthcare-as-a-Service (HaaS) to genetrate during this surroundings. The cloud-based infrastructure as a repository for storage and applying analytical algorithms for retrieval of data regarding the patients has been used in the existence programme. To apply analytics, algorithms for cluster formation and information retrieval area unit designed on the idea of Expectation-

Maximization and fuzzy rule-based classifier. The performance is analysed with reference to varied analysis metrics specifically average time interval, accuracy, computation value, classification time and false positive rate. The results obtained is effective in finding out the probable patients littered with a specific malady. Moreover, it's performed higher compared with its counterparts specifically multilayer, Bayes network and multidimensional language in terms of classification time and false positive rate.

3. PROPSED SYSTEM

With the evolution of big data analytics technology, a lot of attention has been paid to malady prediction from the perspective of big data analysis, varied researches have been conducted by choosing the characteristics mechanically from an outsized range of knowledge the accuracy of risk classification rather than the antecedently chosen characteristics. The big data analytics consists of structured and unstructured data.

STRUCTURED DATA

Structured data usually resides in relational databases (RDBMS). Fields store length-delineated phone numbers, Social indemnity numbers or nada codes. Even text strings of variable length like names are contained in records, creating it a straight forward pertain search. Data may be human or machine generated as long because the data is created within an RDBMS structure. This format is eminently searchable each with human generated queries and via algorithms victimisation of data and field names, such as alphabetical or numeric, currency or date.

Common relational database applications with structured data embody airline reservation systems, internal control, sales transactions, and ATM activity. Structured Query Language (SQL) permits queries on this kind of structured data within relational databases.

Some RDMS do store or purpose to unstructured data like customers relationship management (CRM) applications. The combination can be awkward at best since memo fields don't loan themselves to ancient database queries.

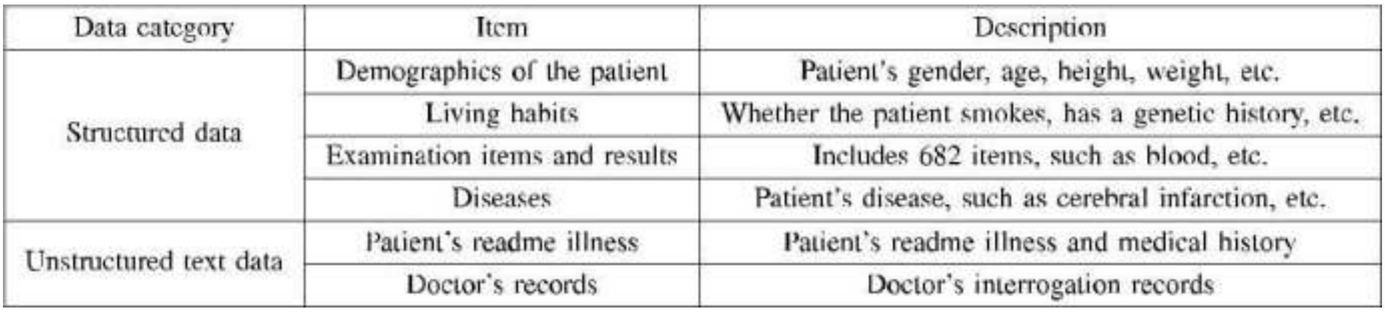
UNSTRUCTURED DATA

Unstructured data is basically everything else. Unstructured data has internal structure however isn't structured via pre-defined data models or schema. It may be in text or non-textual, and human or machine-generated. It may even be kept stored within a non-relational database like NoSQL. Typical human-generated unstructured data includes:

- **Text files:** Data processing, unfold sheets, presentations, email, logs.
- **Email:** Email has some internal structure , and that we typically see it as semi structured. However, its message field is unstructured and ancient analytics tools cannot parse it.
- **Social Media:** Data from Facebook, Twitter, LinkedIn.
- **Website:** YouTube, Instagram, photograph sharing sites.
- **Mobile data:** Text messages and locations.
- **Communications:** Chat, IM, phone recordings, collaboration software.
- **Media:** photographs, audio and video files.
- **Business applications:** MS Office documents, productivity applications.
- **Satellite imagery:** Weather data, land forms, military movements.
- **Scientific data:** Oil and gas exploration, space exploration, seismial imagery, atmospheric data.
- **Digital surveillance:** Surveillance photos and video.
- **Sensor data:** Traffic, weather, oceanographic sensors etc.

In proposed scheme, we combine the structured and unstructured data in healthcare to assess the risk of disease. First, we tend to used latent issue model to reconstruct the missing information from the medical records collected from a hospital in central China. Second, by exploiting statistical knowledge, we could determine the major chronic diseases within the region. Third, to handle structured data, we tend to seek advice from hospital specialists to extract useful features. For unstructured text data, we select the features automatically using CNN algorithm. Finally, we have tendency to propose a unique CNN-based multimodal malady risk prediction (CNN-MDRP) algorithm for structured and unstructured data. The malady risk model is obtained by the combination of structured and unstructured features. Through the experiment, we have a tendency to draw a conclusion that the performance of CNN-MDPR is better than other existing methods.

TABLE 1. CATEGORISATION OF STRUCTURED AND UNSTRUCTURED 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

4. PROCESS OF DATASET AND MODE DESCRIPTION

HOSPITAL DATA

The hospital data set utilized in this study contains real- life hospital data, and the collected data are stored in the data center. To protect the patient's privacy and security, we tend to create a security access mechanism. The data provided by the hospital embody EHR, medical image data and cistron data. We use a three year data set from 2015 to 2018. Our data focus on inmate department data which included 31919 hospitalized patients with 20320848 records in total. The inmate department data is especially composed of structured and unstructured text data. The structured data includes laboratory data and also the patient's basic information like the patient's age, gender and life habits, etc. While the unstructured text information includes the patient's narration of his/her ill health, the doctor's interrogation records and diagnosis, etc. In order to grant out the most malady that have an effect on the region, we have made a statistics on the number of patients, the sex ratio of patients and the major malady in this region every year from the structured and unstructured text data.

DISEASE RISK PREDICTION:

The goal of this study is to predict whether a patient is amongst the cerebral infarct speculative population in line with their medical record. More formally, we regard the risk prediction model for cerebral infraction as the supervised learning methods of machine learning, i.e., the input value is the attribute valu of the patient, X D ($x1$; $x2$; ; xn) 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 patient is amongst the cerebral infarction high-risk population.f$C0$; $C1$g, where, $C0$ indicates the patient is at high-risk of cerebral infarction, $C1$ indicates the patient is at low-risk of cerebral infarct.

For dataset, consistent with the various characteristics of the patient and also the discussion with doctors, we will focus on the subsequent three datasets to achieve a conclusion. Structured data (S-data): use the patient's structured data to predict whether the patient is at high- risk or not of cerebral infarction. Text data (T-data): use the patient's unstructured text data to predict whether the patient is at high-risk or not of cerebral infarction. Structured and text data (S&T-data): use the S-data and T-data on top of multi-dimensionally fuse the structured data and unstructured text data to predict whether the patient is at high-risk or not of cerebral infarction.

EVALUTION METHOD

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 range of instances properly foretold as not required) and false negative (the range of instances incorrectly foretold as not required), respectively.

In addition to the mentioned 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 actual positive rate (TPR) and therefore the false positive rate (FPR), wherever the *TPR* and FPR are outlined as follows:

1. 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 nearer to 1, the model is better. In medical data, we tend to pay additional attention to the recall instead of accuracy. The higher the recall rate, the lower the chance that a patient, who will have the risk of malady is foretold to possess no malady risk.

METHODS FOR DISEASE RISK PREDICTION

DATA IMPUTATION

For patient's examination data, there's an outsized variety of missing data due to human error. Thus, we need to parallel the structured data. Before data imputation, we tend to initial determine unsure or incomplete medical data so modify or delete them to enhance the data quality. Then, we use data integration for data pre-processing. We can integrate the medical data to ensure the data atomicity: i.e., we tend to integrated the height and weight to obtain body mass index (BMI). For data imputation, we use the latent factor model which is presented to explain the observable variables in terms of the latent variables. Accordingly, assume that $Rm\ n$ is the data matrix in our healthcare model. The row designation, m represents the overall range of the patients, and also the column designation, n represents each patient's range of feature attributes. Assuming that k are latent factors, the original matrix R can be approximated as

$$R(m,n)\approx Pm \text{ x } k\ Q^{T}n \text{ x } k \qquad (1)$$

Thus, each element value can be written as $\hat{r}uv = pTu\ qv$, where pu is the vector of the user factor, which indicates the patient's preference to these potential factors, and qv is that the vector of the feature attribute factor. The values of pu and qv are unknown in the above formula.

CNN-BASED UNIMODAL DISEASE RISK PREDICTION (CNN-UDRP) ALGORITHM

For the process of medical text data, we have a tendency to utilize CNN -based unimodal malady risk prediction (CNN-UDRP) algorithm which might be divided into the subsequent five steps.

1) REPRESENTATION OF TEXT DATA

As for every word within the medical text, we have a tendency to use the distributed illustration of Word Embedding in language processing, i.e. the text is delineate within the style of vector. In this experiment, each word will be represented as a Rd -dimensional vector, where d D 50. Thus, a text including n words can be represented as

T D ($t1$; $t2$; ; tn),T 2 $R^{d\ n}$.

2) CONVOLUTION LAYER OF TEXT CNN

We choose two words from the front and back of each word vector $ti0$ in the text, i.e. use the row vector as the representation, to consist a 50×5 = 250 row vector, i.e. $si = (ti^{0}\ 2; ti^{0}\ 1; ti^{0}\ ; ti^{0}C1; ti^{0}C2)$. For $s1$, $s2$, sn-1 and sn, we adopt an zero vector to ll. The selected weight matrix $W1 \in R100\ 250$,i.e., weight matrix $W1$ includes 100 convolution filters and the size of each filter regions is 250. Perform convolution operation on $W1$ and si (i=1,2,..n). Specific calculation progress is that: where i = 1,2, ...100, j = 1,2,.. n. $W1[i]$ is the i-th row of weight matrix is the dot product (a sum over element-wise multiplications), $b1 \in R100$ is a bias term, and f (.) is an activation operate (in this experiment, we have a tendency to use tanh-function as activation function). Thus we can get a 100 n feature graph:

h
$1\ \ i,j = f(W^{1}[i]\cdot sj+b^{1})$ (2)

h
$1 = (h^1 i,j)100\times n$ (3)

3) POOL LAYER OF TEXT CNN

Taking the output of convolution layer as the input of pooling layer, we use the max pooling (1-max pooling) operation i.e., selects the max value of the *n* elements of each row in feature graph matrix. After max pooling, we obtain 100 1 features *h*2. The reason of selecting scoop pooling operation is that the role of each word within the text isn't utterly equal, by most pooling we will select the weather which play key role in the text. In spite of different length of the input training set samples, the text is converted into a fixed length vector after convolution layer and pooling layer, for example, we get 100 features of the text.

4) FULL CONNECTION LAYER OF TEXT CNN

Pooling layer is connected with a fully connected neural net-work. The specific calculation process is that:

$h3 = W3h2\ C\ b3$ (4)

Where $h3$ is the value of the full connection layer, W 3 and $b3$ is the corresponding weights and deviation. **CNN-BASED MULTIMODAL DISEASE RISK PREDICTION (CNN-MDRP) ALGORITHM**

- **TRAINING WORD EMBEDDING**

Word vector training needs pure corpus, the purer the better, that is, it's better to use a professional corpus. In this paper, we extracted the text data of all patients in the hospital from the medical large data centre. After improvement these data, we tend to set them as corpus set. Using ICTACLA word segmentation tool, word2vec tool n-skip gram algorithm trains the word vector, word vector dimension is set to 50, after training we get about 52100 words in the word vector.

- **TRAINING PARAMETERS OF CNN-MDRP**

In CNN-MDRP algorithm, the specific training parameters are *W* 1; *Wnew*3; *b*1; *b*3*new*. We use random gradient methodology to train parameters, and at last reach the risk assessment of whether or not the patient suffers from cerebral infarction. Some advanced features shall be tested in future study, such as fractal dimension, bi-orthogonal wavelet transform etc.

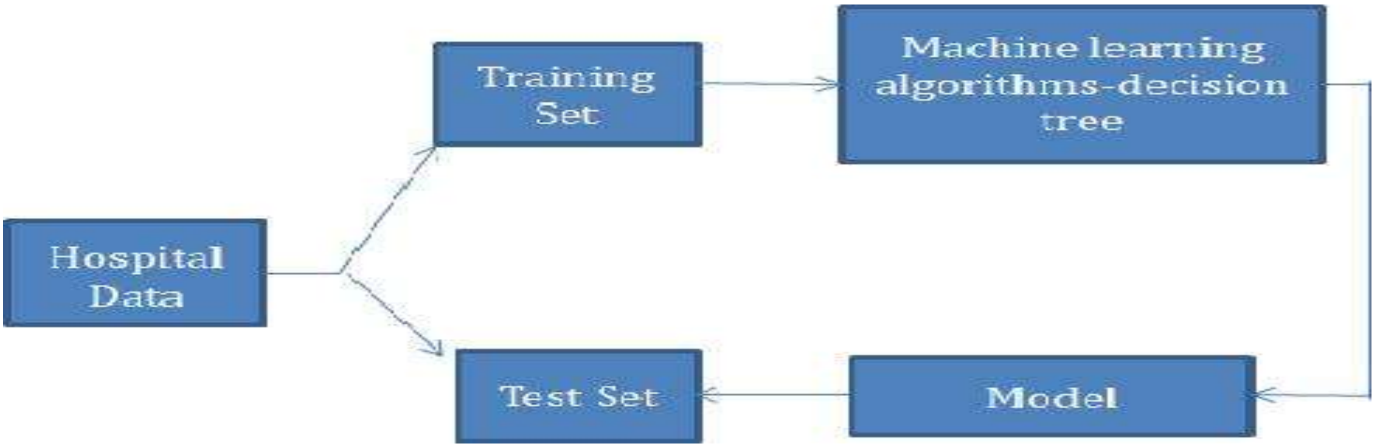


FIG.1: BLOCK DIAGRAM

5. ANALYSIS OF RESULTS

- **STRUCTRED 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 needs to calculate the likelihood of feature attributes. We use contingent probability formula to estimate separate 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's needed to work out the measuring of distance and therefore the choice of k value. In the experiment, the data is normalized rest. Then we have a tendency to use the Euclidean distance to measure the distance.

To determine the most effective and improve the accuracy of the model, the 10-fold cross-validation method is used for the training set, and data from the test set are not used in the training phase. 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 that the feature description of the malady, the higher the accuracy will be. For some simple malady, 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. But for a complex disease, such as cerebral, only using features of structured data is not a good way to describe the disease. The corresponding accuracy is low, that is roughly around 50%. Therefore, we leverage not only the structured data but also the text data of patients based on the proposed CNN-MDPR algorithm. By combining these two data, the accuracy rate can reach HIGH, so as to better evaluate the risk of a disease.

A new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm is done using structured and unstructured data from hospital. To the most effective of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics. Compared to many typical prediction algorithms, the prediction accuracy of our projected algorithm reaches 94.8% with a convergence speed that is quicker than that of the CNN -based unimodal malady risk prediction (CNN-UDRP) algorithm. The model provides the response which can be used for faster decision making. The demand for electricity in a state, trading strategies of the market, patient hospitalization, and patient readmission are the few examples where predictive modelling is used. We can use the supervised learning methods such as classification, regression on the data sets provided. This can derive a model for accurate prediction on various issues. We can iterate the process till we find the best model and then integrate the model into applications to find best predictions.

6. CONCLUSION

Our development and evaluation has shown that collaborative filtering is a strong and viable approach to disease prediction. However, there are still several fascinating avenues for future work. A new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm is done using structured and unstructured data from hospital.

REFERENCE

1) "Research on high-risk student prediction based on big data" Yu Xiaogao, Peng Ruiqing School Of Information Management And Statistics Hubei University Of Economics Wuhan, China -2017.

2) "A Comparative Study On Traditional Healthcare System And Present Healthcare System Using Cloud Computing And Big Data", Manikandan Shanmugam Department Of Computer Science Christ University,2017.

3) "Health care related data in the cloud: challenges and opportunities", Valentina Casola University of Naples

4) "Federico II", Aniello Castiglione University of Salerno ,Kim-Kwang Raymond Choo, University of Texas at San Antonio Christian Esposito University of Salerno,IEEE,2016.

5) "Security-Aware Information Classifications Using Supervised Learning For Cloud-Based Cyber Risk Management In Financial Big Data", Keke Gai1, Meikang Qiu, Sam Adam Elnagdy, IEEE, 2016.

6) P. Groves, B. Kayyali, D. Knott, And S.Van Kuike,"TheBigdata‟Revolution in healthcare:Accelerating value and innovation USA: Center For US Health System" Reform Business Technology Office, 2016.

7) M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey", Mobile Netw. Appl., Vol. 19, No. 2, pp. 171 209,Apr. 2014.

8) P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining Electronic Health Records: Towards Better Research Applications And Clinical Care", Nature Rev.Genet., Vol. 13, No. 6.pp. 395 405, 2012.

9) "An overview of potential factors for effective telemedicine transfer to Sub-Saharan Africa", Peter Meso, Victor W. A. Mbarika, member, IEEE, and Sanjay Prakash Sood, 2009.