A Machine Learning Approach for Disease Prediction

[1]Gireesh Babu C N, [2]Appanna B Prakash, [3]Greeshma Shetty

[1,2,3]BMSIT&M, Bengaluru, Karnataka, India

ABSTRACT -The world is facing problems, such as uneven distribution of medical resources, growth of unpredictable chronic diseases and the increasing medical expenses. Blending the latest information technology into the healthcare system will greatly mitigate the problems. Healthcare data is rapidly growing with the large volume and multi-dimensional data generation from cyber, physical, and social space. Heterogeneous healthcare data in various forms, such as images, text, video, raw sensor data, etc., are required to be effectively stored, processed, queried, indexed and analyzed. These datasets differ widely in their volume, variety, velocity and value, including patient-oriented data such as electronic medical records (EMR), public-oriented data such as public health data, and knowledge-oriented data such as drug-to-drug, drug-to-disease, disease to disease interaction registries. In this survey paper, we combine traditional as well as novel machine learning algorithms for effective prediction of chronic disease outbreak in disease-frequent communities. We survey the modified prediction models over real-life hospital data collected from China over a period of three years. Using big data analytics performed on various biomedical and healthcare stakeholders, accurate analysis of medical data helps with early disease prediction and accordingly helps with efficient healthcare services. But the analysis accuracy reduces when the records are incomplete or the quality of data is missing. A latent factor model is used to reconstruct the incomplete or missing data. The survey is based on predicting the cerebral infarction disease. The algorithm explored in the survey is anew convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm which uses structured and unstructured data from various healthcare stakeholders. The prediction accuracy of the CNN-MDRP algorithm reaches 94.8% compared to other traditional prediction algorithms, with a convergence speed which is faster than that of the CNN-based uni modal disease risk prediction (CNN-UDRP) algorithm.

Keywords *: Big Data Analytics, Healthcare, Machine Learning*

1. INTRODUCTION

Big data analytics is the process of examination of huge data sets which consists of a variety of data types.Big data analytics enables organizations to analyze various kinds of data like structured, semi-structured and unstructured data in order to extract some valuable business information and insights. Machine learning is an application of artificial intelligence (AI). It provides systems the ability to automatically learn and improve from experience without being explicitly programmed. The primary aim is to allow the computers to learn automatically without human intervention or assistance and adjust actions accordingly.

The healthcare system consists of diverse stakeholders such as specialist surgeons, nurses, radiologists, laboratory technologists, pathologists and so on. Each of these stakeholders generate data from heterogeneous sources such as physical examinations, clinical notes, patient interviews and observations, laboratory tests, imaging reports, treatments, therapies, surveys, bills and insurance. Innovative tools and techniques as well as powerful computing technologies are now used to store, process, analyze and extract values from voluminous and heterogeneous healthcare data in a real time manner. Hence, the healthcare system is fast becoming a big data industry. The data acquired from these various healthcare sources can be realized and analyzed using machine learning algorithms to provide information. This information can be related to patients, doctors, diseases, hospitals and so on. Hence the realized information can be used for different purposes depending on the need.

According to research reports [7], 50% of Americans have one or more chronic diseases and 80% of American medical care fee is spent on chronic disease treatment.The healthcare problem of chronic diseases is also very important in many other countries. In China, chronic diseases are the main cause of death and according to a Chinese report on nutrition and chronic diseases in 2015, 86.6% of deaths are caused by chronic diseases. Therefore, it is essential to perform risk assessments for chronic diseases. With the growth in medical data [8], collecting electronic health records (EHR) is increasingly convenient. Proposed models also comprise of healthcare systems which use smart clothing for sustainable health monitoring [6]. Patients' medical information, test results and disease history are recorded in the EHR, which is used to identify potential diagnoses to help save human lives and treat diseases efficiently. Traditional disease prediction risk models usually involve a machine learning algorithm and a supervised learning algorithm by the use of training data with labels to train the model. In the test set, patients can be classified into groups of either high-risk or low-risk [18]. These models are valuable in clinical situations and are widely studied. However, these schemes have a lot of defects. The data set is typically small, for patients and the characteristics for diseases with specific conditions are selected through experience [2]. However, these pre-selected characteristics may not satisfy the changes in the disease and its influencing factors.

With the development of big data analytics, more attention has been paid to disease prediction from the perspective of big data analysis and various researches have been conducted by selecting the characteristics automatically from a large number of data to improve the accuracy of risk classification, rather than the previously selected characteristics. However, those existing work mostly considered structured data. For unstructured data, using a convolutional neural network (CNN) to extract text characteristics has achieved very good results [3[4]. However, there is a large difference between diseases in different regions, primarily because of the diverse climate and living habits in the region.This survey aims to show a model solution that solves these challenges related to disease prediction of cerebral infarction. To solve these problems, the structured and unstructured data in from different healthcare sources are combined to assess the risk of disease. The first step is to use a latent factor modelto reconstruct the missing data from the medical records which is collected from a hospital in central China [14]. Second, by using statistical knowledge, the major chronic diseases in the region are determined. We handle structured data using hospital expertswho are consulted to extract useful features. For unstructured text data, the features are selected automatically using a CNN algorithm.

The proposed CNN-based multimodal disease risk prediction (CNN-MDRP) machine learning algorithm is finally usedto evaluate structured and unstructured data and its performance is evaluated. Through the survey, we draw a conclusion that the performance of CNN-MDPR is better than other existing methods and thus, we combine the results of big data and the feedback obtained from the machine learning algorithms to predict chronic diseases and to be able to treat them before it reaches to the later stages.

2. DATASET AND MODEL DESCRIPTION

A. Data Collection and Classification

The hospital dataset used in the survey consists of both real-life hospital data as well as data stored in the data center. The data provided by the hospital include medical image data,EHR and gene data. For the survey, we use a data set which is collected over a period of three years from 2013 to 2015[14]. Our data focus on inpatient department data which includesaround30 thousand hospitalized patients and a total of precisely 20320848 records. The inpatient department datacontains both structured and unstructured text data. The structured data includes both laboratory data and the patient's basic information such as the patient's age, gender, life habits, etc. Whereas, the unstructured text data includes the patient's narration of his/her illness, the doctors'consultation records and diagnoses, etc.

Table 1 show the real-life hospital data which is first collected and then classified into two categories, namely structured and unstructured text data. In order to retrieve the main disease that affect a region, a statistics is drawn on the number of patients, the sex ratio of patients and the major disease in the region every year from the structured and unstructured text data. The statistical results for the surveyed region are as shown in Table 2. From Table 2, it is observed that the proportion of male and female patients hospitalized each year have little difference and more patients are admitted to the hospital in 2014 [16]. Moreover, the people affected by chronic diseases have always been occupying a large proportion in this area which is shown through the statistics of the data. In our survey, we mainly focus on the risk prediction of cerebral infarction since cerebral infarction is one of the most fatal diseases.

Table 1: Real-life hospital data collected

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 Data	Patient's readme illness	Patient's readme illness and medical history
	Doctor's records	Doctor's interrogation records

Table 2: Statistics From Hospital Data for the period of 3 years

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 Prediction

From Table II, we obtain the main chronic disease in the surveyed region. 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 = (x1; x2; _ _ _; xn) which includes the patient's personal information such as gender,age,living habits and the prevalence of symptoms and other structured data and unstructured data. The output value,*C*, indicates if the patient is amongst the cerebral infarction high-risk population. It is given by:C = {C_0,C_1} where, C_0 indicates the patient is at high risk of cerebral infarction and C_1 indicates the patient is at low risk of cerebral infarction [4]. Hence, the goal of this survey is to predict whether a patient is amongst the cerebral infarction high-risk population according to analysis done on their medical reports. The following introduces the dataset, experiment setting, dataset characteristics and learning algorithms briefly. According to the varied characteristics of the patient and discussions with doctors, we focus on three different types datasets for data collection:

i)Structured data (S-data): The patient's structured data is used to predict whether the patient is at high-risk of cerebral infarction or not.

ii) Text data (T-data): The patient's unstructured text data is used to predict whether the patient is at high-risk of cerebral infarction.

iii) Structured and text data (S&T-data): The S-data and T-data use both the 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 around 700 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. The C++ language is used to implement the machine learning and deep learning algorithms and run it in a parallel fashion by the use of a data center. In our proposed survey, for S-data, we extract the patient's demographics characteristics and some of the characteristics associated with cerebral infarction and living habits. Then, a total of each patient's 79 features are obtained. For T-data, we first extract 815073 words in the text to learn Word Embedding. Then we utilize the independent feature extraction by CNN (Convolutional Neural Network. For S-data, we use Naive Bayesian (NB), K-nearest Neighbour (KNN), and Decision Tree (DT) algorithms to predict the risk of cerebral infarction disease. This is because these three machine learning methods are widely used. For T-data, a CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm to predict the risk of cerebral infarction disease is proposed, where CNN-UDRP (T-data) denotes the CNN-UDRP algorithm used for T-data. The risk of cerebral infarction is predicted using the proposed novel CNN-MDRP algorithm for S&T data, which is denoted by CNN-MDRP(S&T-data) for the sake of simplicity.

C. Design and Framework

The model's basic framework is shown in figure 1 below [14].

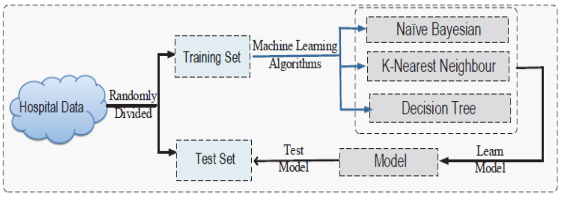

Figure 1: The model's basic framework

For S-data, we use three traditional machine learning algorithms: Naïve Bayesian (NB), K-Nearest Neighbour (KNN) and Decision Tree (DT) algorithm to predict the risk of cerebral infarction. NB classification is a simple probabilistic classifier which is used to calculate the probability of feature attributes. In our survey, 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. We choose classification and regression tree (CART) algorithm among several decision tree (DT) algorithms. In summary, for S-data, the NB classification is the best among all three. However, because cerebral infarction is a disease with complex symptoms, we cannot predict whether the patient is in a high risk group of cerebral infarction only in the light of these simple features. Unstructured data is important to fill the gaps in the structured data. For the patient examination data, there is a large number of missing data due to human errors. Thus, we need to fill the structured data. Before data imputation, we first identify uncertain or incomplete medical data and then modify or delete them to improve the data quality. Then, we use data integration for data pre-processing. For the processing of medical text data, we utilize CNN based unimodal disease risk prediction (CNN-UDRP) algorithm. We find that CNN-UDRP only uses the text data to predict whether the patient is at high risk of cerebral infarction. As for structured and unstructured text data, we survey a CNN-MDRP algorithm based on CNN-UDRP. The accuracy of CNN-MDRP (S&T-data) algorithm is more stable than CNN-UDRP (T-data) algorithm meaning the CNN-MDRP (S&T-data) algorithm reduces error rate after adding structured data. After adding structured data, the recall of CNN-MDRP (S&T-data) algorithm is higher than CNN-UDRP (T-data) algorithm as well.

3. EVALUATION METHODS

For the performance evaluation of the survey 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 [14][16]:

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

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

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

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:

$$\text{TPR} = \frac{TP}{TP + FN}$$

$$\text{FPR} = \frac{FP}{FP + TN}$$

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.

4. RESULTS AND DISCUSSIONS

Figure 2 shows the accuracy, precision, recall and F1-Measure of the S-data using NB, KNN and DT algorithms. We find that the accuracy of the three machine learning algorithms is roughly around 50%. Among them, the accuracy of DT which is 63% is highest, followed by NB and KNN. The recall of NB is 0.80 which is the highest, followed by DT and KNN. We can infer that the corresponding AUC of NB, KNN and DB are 0.4950, 0.4536 and 0.6463, respectively.

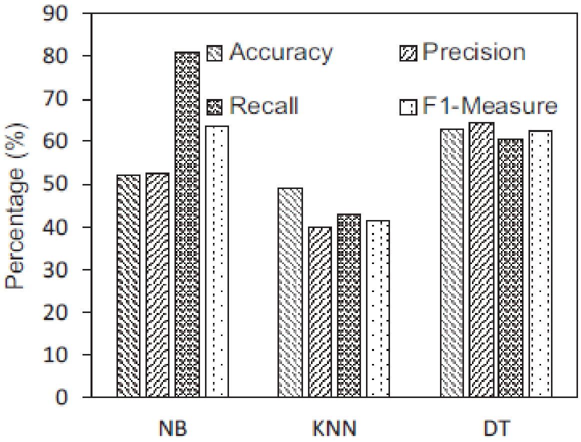

Figure 2: Results of S data[14]

From the figure 3, 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).

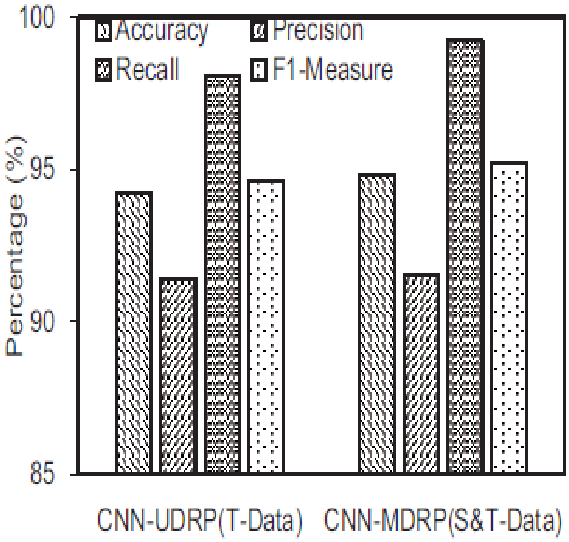

Figure 3: Results of T data and S&T data[14]

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 like hyperlipidaemia, only a few features of structured data can get a good description of the disease, resulting in fairly good effect of disease risk prediction. But for a complex disease, such as cerebral infarction mentioned in the paper, only using features of structured data is not a good way to describe the disease. As seen from Fig. 2 and Fig. 3, the corresponding accuracy is drastically different. Hence, our survey papers use a model that 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.

5. APPLICATIONS

The healthcare environment is generally perceived as being ‘information rich’ yet ‘knowledge poor’. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. The proposed solution integrates big data tools and machine learning algorithms and helps with disease prediction. Our proposed solution, as shown in the experiment, was applied on a data set with limited attributes and the accuracy was around 94%. If sufficient and accurate real time data sets are available then the predictive analysis can be implemented in a better way. HIV, Auto immune diseases, Alzheimer’s, cancer, cerebral infarction itself and many such fatal diseases can be traced at the beginning stages and treated at the earliest. Detection of a disease at an early stage enables patients to have a much higher chance of survival. Therefore, even though we can’t prevent all the diseases from affecting us, we can certainly detect these diseases at an early stage so as to provide the appropriate medical help at the right time to the right people.

6. CONCLUSION

The latest information technologies can be used in the healthcare field to overcome worldwide health problems such as uneven distribution of medical resources, the growing chronic diseases, and the increasing medical expenses. In this survey paper, we explore a proposed solution which is based on big data and machine learning. From this survey paper, a new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm is used to analyze both structured and unstructured data from various healthcare communities. Comparatively, the prediction accuracy of the surveyed algorithm accounts up to 94.8% with a convergence speed that is much faster than that of the existing CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm. The accuracy is hindered a bit as the dataset is not complete. The missing attribute features account for the loss in accuracy. Hence, real time data, when extrapolated to perform the same analytics using the surveyed algorithms, can reach an accuracy of almost a 100 percent. Thus, disease prediction depends on the quality and quantity of data provided by the healthcare communities. Using the surveyed disease prediction model, we can predict diseases before it reaches a point of no return and hence appropriate diagnoses can be performed on patients to help them as far as possible.

REFERENCES

[1] S.-H. Wang, T.-M. Zhan, Y. Chen, Y. Zhang, M. Yang, H.-M. Lu, H.- N. Wang, B. Liu, and P. Phillips, “Multiple sclerosis detection based on biorthogonal wavelet transform, rbf kernel principal component analysis, and logistic regression,” IEEE Access, vol. 4, pp. 7567–7576, 2016.

[2] 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.

[3] J. Wan, S. Tang, D. Li, S. Wang, C. Liu, H. Abbas and A. Vasilakos, “A Manufacturing Big Data Solution for Active Preventive Maintenance”, IEEE Transactions on Industrial Informatics, 2017.

[4] W. Yin and H. Sch¨utze, “Convolutional neural network for paraphrase identification.” in HLT-NAACL, 2015.

[5] S.-M. Chu, W.-T. Shih, Y.-H. Yang, P.-C. Chen, and Y.-H. Chu, “Use of traditional chinese medicine in patients with hyperlipidemia: A population-based study in taiwan,” Journal of ethnopharmacology, 2015

[5] [6] 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, Jan. 2017.

[7] P. Groves, B. Kayyali, D. Knott, and S. V. Kuiken, “The’bigdata’revolution in healthcare: Accelerating value and innovation,” 2016.

[8] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” Mobile Networks and Applications, 2014.

[9] 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.

[10] Y.-D. Zhang, X.-Q. Chen, T.-M. Zhan, Z.-Q. Jiao, Y. Sun, Z.-M. Chen, Y. Yao, L.-T. Fang, Y.-D. Lv, and S.-H. Wang, “Fractal dimension estimation for developing pathological brain detection system based on minkowski-bouligand method,” IEEE Access, vol. 4, pp. 5937–5947, 2016.

[11] J. Wang, M. Qiu, and B. Guo, “Enabling real-time information service on telehealth system over cloud-based big data platform,” Journal of Systems Architecture, 2017.

[12] 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, 2014.

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

[14] Min Chen, YixueHao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang “Disease Prediction by Machine Learning over Big Data from Healthcare Communities”, IEEE Transactions, 2016.

[15] 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.

[16] N. Nori, H. Kashima, K. Yamashita, H. Ikai, and Y. Imanaka, “Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care,” in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.

[17]S. Zhai, K.-h. Chang, R. Zhang, and Z. M. Zhang, “Deepintent: Learning attentions for online advertising with recurrent neural networks,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.

[18] Yujun Ma, (Member, IEEE), Yulei Wang, Jun Yang, YimingMlao, and Wei Li “Big Health Application System based on Health Internet of Things and Big Data”, IEEE, Translations, 2016