Cerebral Infarction Prediction by Machine Learning Over Big Data

Nishmitha K Uchil
Department of CS&E
P. A. C. E., Mangaluru, India

Dr. M Sharmila Kumari
Department of CS&E
P. A. C. E., Mangaluru, India

***Abstract*— With huge growth in biomedical and social insurance networks, precise investigation of therapeutic information benefits early ailment discovery. Nonetheless, the investigation exactness is diminished when the nature of restorative information is inadequate. Additionally, extraordinary locales show novel attributes of certain provincial illnesses, which may debilitate the forecast of sickness episodes. The changed expectation models were tested over genuine medical clinic information gathered from focal China in 2013-2015. To conquer the trouble of deficient information, they have utilized a dormant factor model to reproduce the missing information. The convolutional neural network system based multimodal disease risk prediction (CNN-MDRP) was proposed for calculation and utilizing of organized and unstructured information from emergency clinic. To the best of their insight, none of the current work concentrated on the two information types in the zone of therapeutic huge information examination. Contrasted with a few run of the mill forecast calculations, the expectation precision of their proposed calculation achieves 94.8% with an intermingling speed which is quicker than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) calculation.**

I. INTRODUCTION

Human face lots of problems related to the chronic disease. The main reason behind increase the chronic disease such as improper living habits, insufficient physical exercise, unhealthy diet, and irregular sleeping [3]. 80% of people in the United States, spent more amount on the diagnosis of chronic disease [1]. People give more aid for accurate prediction of disease [1]. In many regions, different diseases cause due to the environmental factors and lifestyle of people [1]. Sometimes it may lead to the wrong decision making regarding the disease prediction. Due to preliminary disease prediction, it can reduce the risk of disease and patient gets diagnosed as early as possible. For the prediction of disease with the help of IoT device is done so the data collection used sensor generated data but the sensors are uncomfortable to the user and required multiple sensors to wear [3].In the previous year, data mining plays an important role in the healthcare system. In this KDD process involve, extract undiscoverable knowledge with the help of target data [7]. Data mining divided into two part, first one is predictive and the second one is descriptive [7].The predictive part consists of classification and regression, whereas descriptive part consists of clustering and association rule [7]. Data mining is the way of disease prediction depends on historical records of patients. Previous work mostly based on disease prediction who is automatically extracting a large number of features from data for better accuracy of the system [7]. Structured data is widely used for the disease prediction other than unstructured data. But by the use of a convolutional neural network, it becomes easy to deal with unstructured data also [1]. The convolutional neural network is deep learning algorithm that extracts the features automatically from the large dataset and gets the proper result [1].

II. OBJECTIVE

The analysis accuracy is reduced when the quality of medical data in incomplete. Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks. However, those existing work mostly considered structured data. There is no proper methods to handle semi structured and unstructured. The proposed system will consider both structured and unstructured data. The analysis accuracy is increased by using Machine Learning algorithm and Map Reduce algorithm.

III. 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-life hospital data, and the data are stored in the data center. To protect the patient's privacy and security, they created a security access mechanism. They used three year data set from 2013 to 2015. Their data focus on inpatient department data which included 31919 hospitalized patients with 20320848 records in total.

B. Disease Risk Prediction

For dataset, according to the different characteristics of the patient and the discussion with doctors, they have focused 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.

C. Evaluation Methods

For the performance evaluation in the experiment. First, we denote

- TP - TRUE POSITIVE - Number of instances correctly predicted as required.
- FP - FALSE POSITIVE - Number of instances incorrectly predicted as required.
- TN - TRUE NEGATIVE - Number of instances correctly predicted as not required.
- FN - FALSE NEGATIVE - Number of instances incorrectly predicted as not required.

Then, we can obtain four measurements: accuracy, precision, recall and F1-measure as follows:

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

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

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

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

IV. METHODS

This section includes data imputation, CNN based unimodal disease risk prediction (CNN-UDRP) algorithm and CNN-based multimodal disease risk prediction (CNN-MDRP) algorithm.

A. CNN- based Multimodal Disease Risk Prediction(CNN-MDRP)

For structured and unstructured text data, CNN-MDRP algorithm is designed as shown in Fig. 1. As shown in Fig. 1(a-d), which can extract 100 features about text data set. For structure data, they extracted 79 features. Then, they conducted the feature level fusion by using 79 features in the S-data and 100 features in T-data. Since the variation of features number, the corresponding weight matrix and bias change to W3 new, b3 new, respectively. They also utilized softmax classifier. In the following they will introduce how to train the CNN-MDRP algorithm, the specific training process is divided into two parts.

1) Training and Embedding: Word vector training requires pure corpus, the purer the better, that is, it is better to use a professional corpus. In this paper, they extracted the text data of all patients in the hospital from the medical large data center. After cleaning these data, we set them as corpus set. Using ICTACLAS 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.

2) Training parameters of CNN-MDRP: In CNN-MDRP algorithm, the specific training parameters are W^1 , W^3 new, b^1 , b^3 new. They used stochastic gradient method to train parameters, and finally reach the risk assessment of whether the patient suffers from cerebral infarction. Some advanced features shall be tested in future study, such as fractal dimension , biorthogonal wavelet transform etc.

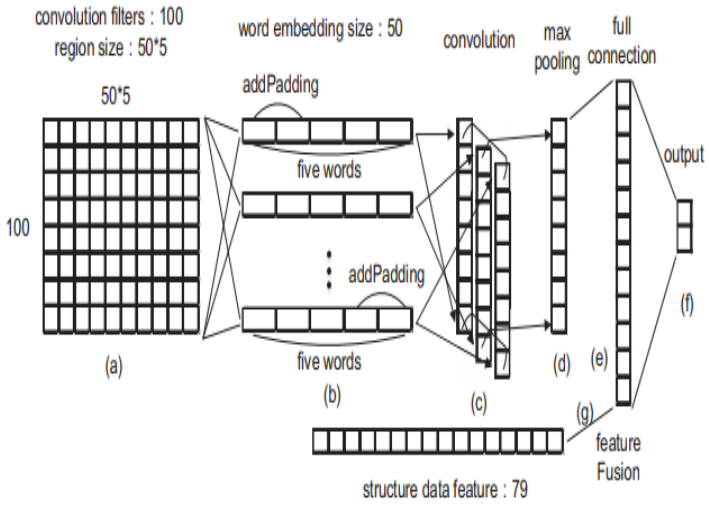

Fig 1-CNN-based multimodal disease risk prediction (CNN-MDRP) algorithm.

V. RESULT

This section, describes the overall results about S-data and S&T-data.

A. Structured Data (S-data): For S-data, they used 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, they used 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 they used the Euclidean distance to measure the distance. As for the selection of parameters k, they found that the model is the best when k = 10. Thus k = 10. The choose classification and regression tree (CART) algorithm among several decision tree (DT) algorithms. To determine the best classifier 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. We can see that the accuracy of the three machine learning algorithms are 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. Corresponding AUC of NB, KNN and DB are 0.4950, 0.4536 and 0.6463, respectively. In summary, for S-data, the NB classification is the best in experiment. However, it is also observed that we cannot accurately predict whether the patient is in a high risk of cerebral infarction according to the patient's age, gender, clinical laboratory and other structured data. In other word, because cerebral infarction is a disease with complex symptom, they cannot predict whether the patient is in a high risk group of cerebral infarction only in the light of these simple features.

B. Structured and Text Data (S&T-data): According to the discussion before, they got the accuracy, precision, recall, F1-measure and ROC curve under CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms. 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, they 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, they 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 CNNUDRP (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. 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. The corresponding accuracy is low, which is roughly around 50%. Therefore, in this paper, they leverage not only the structured data but also the text data of patients based on the proposed CNN-MDPR algorithm. They found that by combining these two data, the accuracy rate can reach 94.80%, so as to better evaluate the risk of cerebral infarction disease

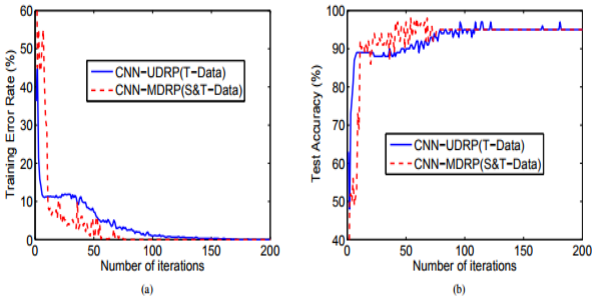

Figure 2.1 : Effect of iterations on the algorithm. (a) The trend of training error rate with the iterations for CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms. (b) The trend of test accuracy with the iterations for CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms.

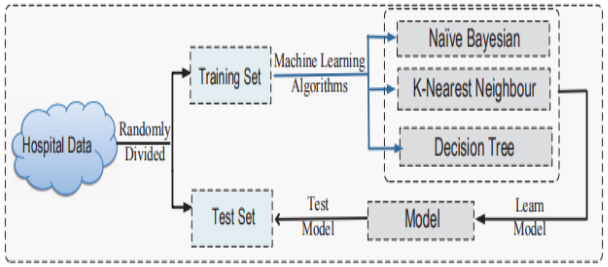

Figure 2.2: The three machine learning algorithms used in the disease prediction experiments.

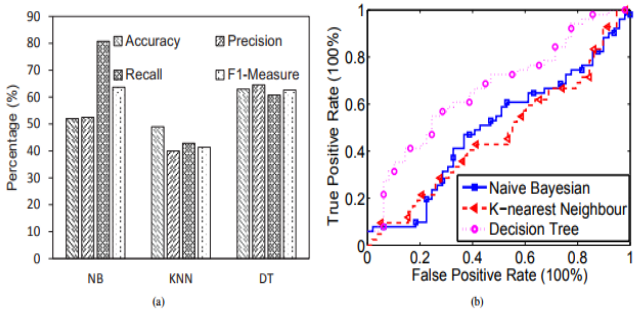

Figure 2.3: Overall results of S-data. (a) Comparison of accuracy, precision, recall and F1-Measure under S-data for NB, KNN and DT, in which NB = naive Bayesian, KNN = k-nearest neighbor, and DT = decision tree. (b) ROC curves under S-data for NB, KNN and DT.

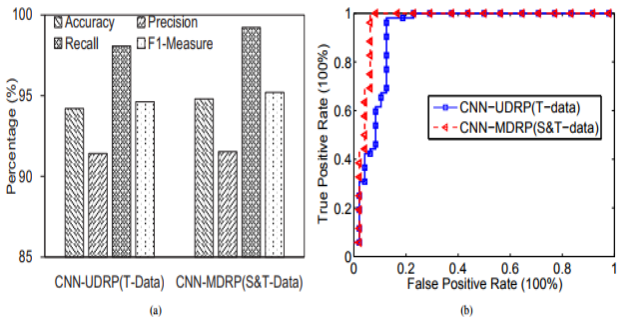

Figure 5.4: Overall results of S-data. (a) Comparison of accuracy, precision, recall and F1-Measure under S-data for NB, KNN and DT, in which NB = naive Bayesian, KNN = k-nearest neighbor, and DT = decision tree. (b) ROC curves under S-data for NB, KNN and DT.

CONCLUSION

In this paper, convolutional neural system based multimodal infection hazard expectation calculation utilizing organized and unstructured information from emergency clinic is proposed. To the best of their insight, none of the current work concentrated on the two information types in the territory of restorative enormous information investigation. Contrasted with a few run of the mill forecast calculations, the expectation exactness of their proposed calculation achieves 94.8% with a union speed which is quicker than that of the CNN-based unimodal sickness chance expectation calculation.

REFERENCES

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, ‘‘Disease prediction by machine learning over big data from healthcare communities,’’ IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.

[2] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang,‘‘ A relative similarity basedmethodforinteractivepatientriskprediction,’’DataMiningKnowl.Discovery, vol. 29, no. 4, pp. 1070– 1093, 2015.

[3] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, ‘‘Wearable 2.0: Enable human-cloud integration in next generation healthcare system,’’ IEEE Commun. , vol. 55, no. 1, pp. 54–61, Jan. 2017.

[4] 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 Syst. J., vol. 11, no. 1, pp. 88–95, Mar. 2017.

[5] L. Qiu, K. Gai, and M. Qiu, ‘‘Optimal big data sharing approach for telehealth in cloud computing, ’’in Proc. IEEE Int. Conf. Smart Cloud (Smart Cloud), Nov. 2016, pp. 184–189.

[6] Disease and symptoms Dataset -WWW.Github.com. [7]Heart disease Dataset-WWW.UCI Repository.com. [8]Ajinkya Kunjir, Harshal Sawant, Nuzhat F.Shaikh, "Data Mining and Visualization for prediction of Multiple Diseases in Healthcare" in IEEE big data analytics and computational intelligence, Oct 2017 pp.23-25.