RESEARCH ARTICLE OPEN ACCESS

A SURVEY ON DISEASE PREDICTION IN BIG DATA HEALTHCARE USING EXTENDED CONVOLUTIONAL NEURAL NETWORK

Asadi Srinivasulu, S.AmruthaValli, P.Hussainkhan, and P.Anitha
Department of Information Technology, Sree Vidyanikethan Engineering College, A.Rangampet
Email: srinu.asadi@gmail.com, srikakulaamruthavalli@gmail.com, patanhussainkhan@gmail.com, anitharajpatel0110@gmail.com

Abstract: Different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks. With big data growth in biomedical and healthcare communities, accurate analysis of medical data benefits early disease detection, patient care, and community services. So machine learning algorithms for effective prediction of chronic disease outbreak in disease-frequent communities are adopted. But analysis accuracy is reduced when the quality of medical data is incomplete. A new convolutional neural network (CNN)-based risk prediction algorithm using structured and unstructured data from hospital is implemented. Compared with several typical prediction algorithms, the proposed system produces high accuracy, high performance, and high convergence speed.

***Keywords* — Big data analytics, machine learning, health care, convolutional neural network (CNN), Prediction algorithm.**

I. INTRODUCTION

First determine the major chronic diseases in the region. Next to handle structured data, consult with hospital experts to extract useful features. For unstructured text data, selecting the features automatically using CNN algorithm. Finally, a novel CNN-based algorithm for structured and unstructured data is proposed.The disease risk model is obtained by the combination of both structured and unstructured features.The characteristics are selected through experience. However, these pre-selected characteristics maybe not satisfy the changes in the disease and its influencing factors.

With the development of big data analytics technology, more attention has been paid to disease prediction from the perspective of big data analysis, 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.

In this paper, we mainly focus on the risk prediction of liver disease. The Hospital dataset related to liver disease in structured data format. The goal of this study is to predict whether a patient is amongst high-risk population according to their medical history. More formally, we regard the risk prediction model for as the supervised learning methods of machine learning, i.e., the input value is the attribute value of the patient, $X = (x_1, x_2, \cdots, x_n)$ which includes the patient's personal information such as age, gender and other structured data as mentioned above .

The output value is C, which indicates whether the patient is amongst the high-risk population. C ={C0,C1}, where, C0 indicates the patient is at high-risk C1 indicates the patient is at low-risk. The following will introduce the dataset, experiment setting, dataset characteristics and learning algorithms briefly.

In the experiment setting and dataset characteristics, we select 583 patients in total as the experiment data and randomly divided the data into training data and testdata.390 patients as the training data set while 193 patients as the test data set. We use the python language .In this paper, for S-data, according to the discussion with doctors and with correlation analysis, we extract the patient's demographics characteristics and some of the characteristics associated with liver disease.

We will introduce machine learning used in this work briefly. For S-data, we use conventional machine learning algorithms, i.e., Naive Bayesian (NB) and Decision Tree (DT) algorithm to predict the risk of disease. This is because these machine learning methods are widely used. For text data use the patient's unstructured text data on which convolutional neural network techniques used to predict whether the patient is at high-risk.

II. METHODOLOGY

For handling structured data we use Decision Tree and Naive Bayies to predict risk. For handling text data we use Convolutional neural network techniques to predict the risk.

A.)DECISION TREE:

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node.

Given a tuple, X, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple.

Algorithm:

Generate.decision.tree.Generate a decision tree from the training tuples of data partition, D

Input:

- Data partition, D, which is a set of training tuples and their associated class labels;
- Attribute, list, the set of candidate attributes;
- Attribute selection. Method, a procedure to determine the splitting criterion consists of splitting attribute and, possibly, either a split-point or splitting subset.

Output: A decision tree.

Method:

1. Create a node N;
2. If tuples in D are al of the same class C, then
2.1 return N as a leaf node labelled with class C;
3. If attribute. List is empty then
3.1 return N as a leaf node labelled with the majority class in D;
4. Apply Attribute_selection_method(D,attribute.list) to find the "best" splitting criterion ;
5. Label node N with splitting-criterion;
6. If splitting_attribute is discrete-valued and multiway splits allowed then
6.1attribute_list attribute_list splitting_attribute,
7. for each outcome of j of splitting.criterion
7.1 Let Dj be the set of data tuples in D satisfying outcome j;
7.2 if Dj is empty then
7.2.1attach a leaf labeled with the majority class in D to node N;
7.3 else
7.3.1attach the node returned by Generate.Decision.tree (Dj,attribute.list) to node N;
8. End for
9. Return N;

Decision trees can easily be converted to classification rules. The construction of decision tree classifiers does not require any domain knowledge or parameter setting.

In general, decision tree classifiers have good accuracy.

B.)NAIVE BAYIES:

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. Given a class variable y and a dependent feature vector x_1 through x_n, Bayes' theorem states the following relationship:

$$P(y \mid x_1 \ldots . x_n) = \frac{P(y)P(x_1 \ldots . x_n \mid y)}{P(x_1 \ldots . x_n)}$$

Using the naive independence assumption that $P(x_i|y, x_1, \ldots, x_{i-1}, x_{i-1}, \ldots, x_n) = P(x_i|y)$. for all i, this relationship is simplified to

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Since $P(x_1, \ldots, x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \ldots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_{y} P(y)\prod_{i=1}^{n} P(x_i \mid y).$$

and we can use Maximum a Posteriori (MAP) estimation to estimate $P(y)$ and $P(x_i \mid y)$; the former is then the relative frequency of class y in the training set. The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P(x_i \mid y)$. Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality.

C.)CONVOLUTIONAL NEURAL NETWORK

This is done in 5 steps as follows:

STEP 1: Representation of Text Data:

- ✓ Word Embedding Model from NaturalLanguage Processing.
- ✓ Each word can be represented as a vector of numerical values (A column matrix).
- ✓ In this experiment, each word will be represented as a **R^d -dimensional vector**, where **d = 50** i.e, we represent each word as a column vector (column matrix) containing 50 rows.
- ✓ Now the text can be represented by appending the column vectors meaning we stack up the column vectors side by side to create of matrix of dimension **dxn.** (Just words are stacked side by side in a sentence).

STEP 2: Convolution Layer of Text CNN

- ✓ Consider there is a text "Patient is at high risk."
- ✓ Each word is represented by 50 column vector.
- ✓ Repeat the following steps until the end of text.(Text contains n words)

1. Start a pointer at position 1.(1st word)
2. Assuming the pointer is at position i, take words at positions i-2 , i-1 , i , i+1, i+2.
3. Transpose each of them to form row matrices of 50 columns and append them side by side converting them into a single row vector of size 50 * 5.
4. Increment the pointer and switch to new row.
5. For first, second and n-1 and nth words we have gaps i.e for the first word we don't have two previous words. In such a case fill them with zero row vectors.
6. At the end of the above process we obtain a nx250 matrix which is our convoluted matrix.
7. The weight matrix $W^1 \in R^{100\times250}$ is of size 100x250. Meaning we are expecting the neural network to extract 100 features for us.
8. Now we carry out the following calculation.

$$\mathbf{h^1_{i,j} = f(W_1[i] \cdot s_j + b_1)}$$

This is the dot product of matrices. b_1 is column matrix of 100 rows. Bias is used to used to shift the

learning process. Without adding it, it is simple weighted sum of features and there is no learning process. We obtain a 100xn feature graph h^1 .f is an activation function which is used to obtain non-linearity. We used tanh activation function.

$$h^1 = (h^1_{i,j})_{100\times n}$$

STEP 3: POOL layer of Text CNN

From the feature graph h1 which is 100xn dimensional, we pick the maximum element in each row of the matrix obtaining 100 maximum values from each row.

From these 100 values we construct a 100x1 matrix h2 (column vector).

The reason of choosing max pooling operation is that the role of every word in the text is not completely equal, by maximum pooling we can choose the elements which play key role in the text.

By the end of Step 3 we have extracted 100 features from unstructured data

$$h^2_j = \max_{1 <= i <= n} h^1_{i,j} \quad j=1,2,---100$$

STEP 4: FULL CONNECTION layer of Text CNN

Then provide this matrix as input to a neural network which carries the following computation which is similar to that of in step 2.(dot product of matrices).W3 is the weight matrix of full connection layer and b3 is bias.

$$h3 = W3h2 + b3$$

STEP 5:CNN Classifier

- ✓ **Softmax** classifier as output classifier which predicts the risk of the disease(high or low).

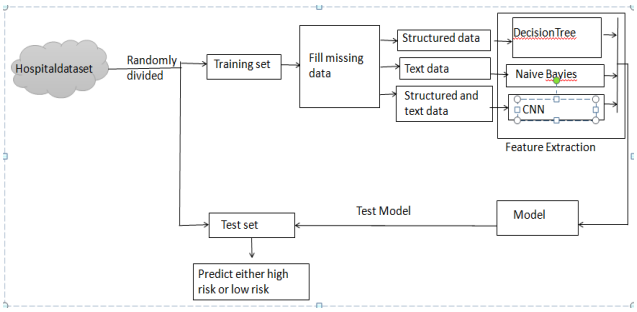

Fig1: Proposed Architecture

III. RESULTS

From the results obtained the accuracy, recall, precision, F1-measure scores are calculated. The accuracy of Decision Tree: 0.663212435233 %.The accuracy of Gaussian Naive Bayes: 0.740932642487 %.The precision of Decision Tree: 0.796875 %.The precision of Gaussian Naive Bayes: 0.779141104294 %.The recall of Decision Tree: 0.723404255319 %.The recall of Gaussian Naive Bayes:0.900709219858 %.The F1-score of Decision Tree: 0.758364312268 %
The F1-score of Gaussian Naive Bayes: 0.835526315789 %.

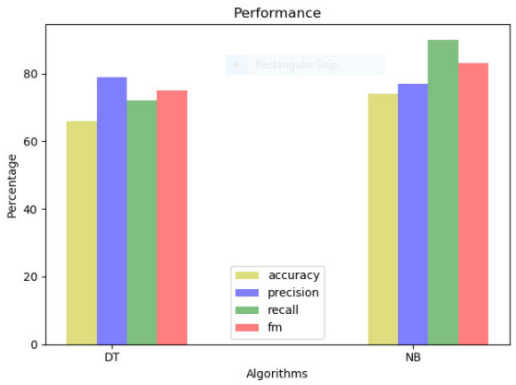

Fig 2: Performance analysis of Decision Tree and Naive Baye sian

Then coming to text data analysis we got 100 features with more performance and accuracy using CNN techniques compared to other algorithms mentioned here. 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.

IV. CONCLUSION

In this paper, we propose machine learning algorithms like Decision Tree and Naive Bayies using structured data from hospital to predict the severity of risk and Convolutional neural network techniques using unstructured data to predict the risk. To the best of our knowledge, 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 reaches high accuracy in predicting the risk.

REFERENCES

1. *K.Karthika, G.Nagarajan, "Disease Prediction By Machine Learning Over Big Data From Healthcare Communities.", Volume 4 Issue 11 Nov 2017.*
2. *W. Yin and H. Sch¨utze, "Convolutional neural network for paraphrase identification." in HLT-NAACL, 2015, pp. 901–911*
3. *P. Groves, B. Kayyali, D. Knott, and S. van Kuiken, The'BigData'Revolution in Healthcare: Accelerating Value and Innovation. USA: Center for US Health System Reform Business Technology Office, 2016.*
4. *M. Chen, S. Mao, and Y. Liu, ''Big data: A survey,'' Mobile Netw. Appl.,vol. 19, no. 2, pp. 171–209, Apr. 2014.*
5. *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.*
6. *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 Trans. Intell. Transp. Syst., vol. 16, no. 6, pp. 3033–3049, Dec. 2015.*
7. *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.*
8. *M. Chen, P. Zhou, and G. Fortino, ''Emotion communication system,'' IEEE Access, vol. 5, pp. 326–337, 2017, doi: 10.1109/ACCESS.2016.2641480.*
9. *M. Chen, Y. Ma, J. Song, C. Lai, and B. Hu, ''Smart clothing: Connecting human with clouds and big data for sustainable health monitoring,'' ACM/Springer Mobile Netw. Appl., vol. 21, no. 5, pp. 825–845, 2016.*
10. *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, vol. 33, no. 7, pp. 1123–1131, 2014.*