A Deep-Learning Based Ultrasound Text Classifier for Predicting Benign and Malignant Thyroid Nodules

Dehua Chen, Jinxuan Niu, Qiao Pan, Yue Li, Mei Wang
School of Computer Science and Technology
Donghua University
Shanghai, China
e-mail:chendehua@dhu.edu.cn

***Abstract*: The diagnosis of benign and malignant thyroid nodules timely and correctly is always a core problem for the clinical practice of thyroid nodules. Traditionally, preoperative diagnosis process of benign and malignant thyroid includes two stages: ultrasound check and fine needle aspiration. The malignant thyroid nodules will be further confirmed by surgery and pathology. However, such traditional diagnosis process falls into trouble in clinical practice since fine needle aspiration has potential risk of turning benign nodules into malignant ones. Moreover, the correct diagnosis of malignant thyroid nodules is to a great extent determined by the expertise of clinicians. As a new machine learning method, deep learning has been applied in computer-aid diagnosis recently. Thus, in this paper, we propose a deep-learning based ultrasound text classifier for predicting benign and malignant thyroid nodules. The proposed ultrasound text classifier is a kind of supervised classification based on deep neural network which is trained by the labeled ultrasonic text with benign or malignant label of pathology. Experimental results show that this method has the highest accuracy rate of 93% and 95% both on the real medical dataset and the UCI standard dataset, compared with the traditional Random Forest, Support Vector Machine and Neural Network both on the real medical dataset and the UCI standard dataset.**

Keywords—Thyroid Nodules; Ultrasound features; Differential Diagnosis; Deep Neural Network

I. INTRODUCTION

Thyroid Nodule is a common disease and the incidence rate is increasing year by year. The preoperative diagnosis of benign and malignant thyroid nodules is a significant step for the clinical treatment of thyroid nodules. Ultrasound check and fine needle aspiration are two traditional kind of preoperative diagnosis methods. After the preoperative diagnosis, the suspected malignant thyroid nodules will be further confirmed by surgery and pathology. However, there are many evidences in clinical practices show that fine needle aspiration has some potential risk of turning benign nodules into malignant ones. In contrast to fine needle aspiration, ultrasound examination is noninvasive to thyroid nodules and easy to operation. Therefore, ultrasound has been widely used in thyroid nodules examination, which led to the creation of large repositories of unstructured ultrasound text in the electronic medical records (EMR).

In the past two decades, several computer-aided diagnosis prediction models of thyroid nodule are developed. With the aid of these models, clinicians are able to identify for patients correctly, and then take the appropriate surgical treatment for malignant nodules in time. However, these existing models are largely based on ultrasonic imaging. The features of ultrasound imaging are a kind of low level ones, which results in low accuracy of prediction of Benign and Malignant Thyroid Nodules.

Compared with ultrasonic imaging, ultrasound text report of thyroid is a detailed description of the sonographic features of thyroid nodules. There are different characteristics in terms of thyroid ultrasound, doctors describe different terms, different diagnostic criteria standardized, diagnosis results depend on the experience of doctors, level, status and other factors, different doctors may give different diagnosis results. Therefore, how to make full use of a large number of thyroid nodules ultrasound text report unstructured data, to build a high accuracy of breast tumor diagnosis model, which is of great practical significance to realize the low cost and safe early prediction of thyroid cancer, improve the quality of thyroid nodules diagnosis and reduce the mortality of thyroid nodules. On the other hand, the correct diagnosis of malignant thyroid nodules is to a great extent determined by the expertise of clinicians. How to leverage the ultrasound text in automatic prediction of benign and malignant thyroid nodules has a potential to improve biomedical research and the delivery of healthcare.

In this paper, we propose a method for the differential diagnosis of thyroid nodules with ultrasound features based on Deep Neural Network (DNN), which can handle the

978-1-5386-2280-3/17 $31.00 © 2017 IEEE
DOI 10.1109/ICGI.2017.39

sparseness and imbalance of training set appropriately, to improve the diagnostic accuracy on text features. Neural network, as one of the most widely used machine learning algorithms, attempts to process information by simulating the network of human brain. It is mainly used to solve the problem of classification and prediction. Because of the multiple parameters and complicated structure, neural network's training effect is better than other machine learning algorithms in most cases. We compare DNN with some traditional machine learning algorithms both on real medical dataset and UCI standard dataset, and the result shows that DNN has the highest accuracy rate with 93% and 95%.

The remainder of the paper is organized as follows: Section 2 reviews related works of differential diagnosis of thyroid nodules. Section 3 introduces how we process data for our experiments. Section 4 describes DNN on model derivation and description. Section 5 presents the experimental results. Section 6 concludes the paper.

II. Related Work

It has become an important work to improve the diagnosis accuracy of diseases through machine learning algorithms. Many scholars have done related researches on differential diagnosis of thyroid nodules.

To solve the sparseness and imbalance problem of dataset, Ma H[1] proposes an effective missing data prediction algorithm, in which information of both users and items is taken into account, and empirical studies on dataset MovieLens show that this method outperforms other state-ofthe-art collaborative filtering algorithms and it is more robust against data sparsity. Saif H[2] uses two different sets of features to alleviate the data sparseness problem on the Stanford Twitter Sentiment Dataset and achieves 86.3% sentiment classification accuracy, which outperforms existing approaches. David Masko[3] uses oversampling technique on the imbalanced training sets to increase the performances of the balanced set and the results show that oversampling is a viable way to counter the impact of imbalances in the training data. Rok Blagus[4] uses three types of sampling techniques to reduce the class-imbalance problem and proves that oversampling techniques unjustifiably appear to perform better than undersampling techniques.

On the study of differential diagnosis of thyroid nodules, Jieming Ma[5] presents a non-invasive and automatic approach for differentiating benign and malignant thyroid nodules with ultrasound elastography based on support vector machines (SVM) with biased penalties, and the results show that this method is able to get maximum geometric mean (MGM) of
90.1% with the sensitivity of 93.8% and the specificity of 86.6%. Young Hun Lee[6] assesses the accuracy of US diagnosis for benign and malignant solid thyroid nodules using a real-time US performance and classification system , which can apparently improve diagnostic accuracy of thyroid US for solid thyroid nodules. Kim E T[7] uses logistic regression to evaluate the computer-aided diagnosis (CAD) of US elastography for classification of benign and malignant thyroid nodules, and the specificity is 83%. Rago T[8] studies 92 consecutive patients with a single thyroid nodule who underwent surgery for compressive symptoms or suspicion of malignancy on fine needle aspiration cytology and illustrates that US elastography has great potential as an adjunctive tool for the differential diagnosis of nodule cancer, especially in indeterminate nodules on cytology.

It can be seen that current researches on classification of thyroid nodules have high requirements on datasets but there is improvement space in prediction accuracy still. DNN algorithm used in this paper not only has the advantage of common neural network algorithm for data compatibility but also improves the accuracy of classification greatly by using more complicated network structure.

III. Data Preprocessing

According to the clinical experience of patients with thyroid nodule and the characteristics of actual ultrasonic indexes data provided by a hospital, we structure the unstructured medical data first, then the feature was carried out, finally we use DNN to train the structured data.

In this paper, we use a structured analisys approach based on dependency syntax analysis. First of all, for the pathological records that often appear in a different description, we use neural network for model training to find the corresponding word vector. After that we find synonyms according to the cosine similarity to uniform specification of the pathological records. Also, sentence segmentation method and word information annotation method are used to simplify the sentence structure, which can reduce the height of the dependency tree maximally either. Besides, it can improve the clarity of the grammatical relationship and make the result of structure more accurate. Then, we use dependency parse method to obtain the dependency relation tree of each phrase, thus the index and corresponding values are extracted automatically according to the semantic feature and speech feature. That is, unstructured text data is transformed into keyvalue structure as a process template. In the end, marked information was restored and noise data get corrected[9].

At the aim of reducing the distortion caused by overlapping information, we use principal component analysis (PCA) to select features to remove unnecessary features under the premise of maximizing the preservation of original data, which can significantly improve both the speed and accuracy of model training at the same time. To do that, the original data should be normalized, corresponding correlation coefficient matrix is calculated and the eigenvalues and eigenvectors are obtained at first. Then we select those principal components corresponding to eigenvalues whose cumulative contribution rate is higher than 85% to replace the original features[10]. As a result, the blood flow pattern, blood supply, echo distribution, envelope contact area, blood supply of thyroid parenchyma, tubercle

boundary and so on are selected as input variables after PCA, and the results of pathologic diagnosis of nodules are output variable on the contrary.

IV. Deep Neural Network

The network structure of the deep neural network is shown in Fig.1. The Network can effectively learn the essence of changeable function from a few samples. The neural network model consists of input layer, hidden layer and output layer, and each layer contains several nodes. There are weights for connection between layers and nodes, nodes and nodes, and the size of weights means the importance of connection. Generally speaking, BP neural network is the most used model. After the input signals delivered from the hidden layer to the output layer, backward propagation is carried out according to the BP algorithm, and the value of weights are adjusted by the gradient descent method at the same time so that the final output is maximized close to the actual expectation[11]. In other words, DNN is a neural network that has several layers.

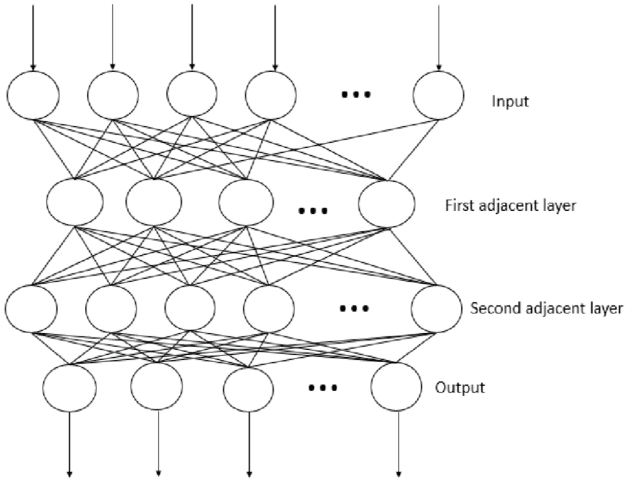

Fig. 1. The network structure of the deep neural network

A. Model Derivation

Let i and j represents the nodes of two adjacent layers respectively, and j is an output layer node, ω_{ji} denotes the connection weight of node i to j, then the input of node j is

$$X_j = \sum_{i=0}^{n} \omega_{ji} y_i \quad (1)$$

Where n denotes the number of nodes at the layer of node i, and y_i denotes the output of the node i. then the output of node j is

$$Y_j = \varphi(X_j) \quad (2)$$

The function φ represents the activation function of the model, the Sigmoid function is selected there.

$$\varphi(x) = \frac{1}{1+e^{-x}} \quad (3)$$

Information from input layer delivered to output layer through formula (1) and (2), then backward propagation is carried out according to the BP algorithm. Assume that the error is e and the true value is d, then

$$e_j = d_j - y_j \quad (4)$$

So the error function is defined as

$$E = \frac{1}{2}\sum_{j\in A} e_j^2 \quad (5)$$

A represents all the nodes of the layer, we can see that E is a function of weight, substituting it to above formulas

$$\frac{\partial E}{\partial \omega_{ji}} = -e_j \varphi_j'(X_j) y_i \quad (6)$$

The change of connection weights between node i and j can be calculated by the gradient descent method

$$\Delta\omega_{ji} = -\mu \frac{\partial E}{\partial \omega_{ji}} = \mu e_j \varphi_j'(X_j) y_i \quad (7)$$

μ represents the learning rate of the model, which can be used to control the magnitude of the weight change. Similarly, suppose h is a node on the hidden layer adjacent to i, then

$$\frac{\partial E}{\partial \omega_{ih}} = \frac{\partial E}{\partial y_i} \varphi_i'(X_i) y_h \quad (8)$$

Considering that the output of the layer at node i would affect the layer where node j is located, then

$$\frac{\partial E}{\partial y_i} = \sum_{j\in A} \frac{\partial E}{\partial y_j} \varphi_j'(X_j) \omega_{ji} \quad (9)$$

In this case, the change of the connection weight between node h and i is

$$\Delta\omega_{ih} = -\mu \frac{\partial E}{\partial \omega_{ih}} = -\mu \left(\sum_{j\in A} \frac{\partial E}{\partial y_j} \varphi_j'(X_j) \omega_{ji}\right) \varphi_i'(X_i) y_h \quad (10)$$

Repeat above training steps until the final error is less than a certain threshold we set before. We illustrate this model in Fig.1.

B. Model Description

Input: training set X, number of hidden layer M, number of nodes per layer N, global error threshold ε, number of learning
K

Output: Instance $i \in$ X belongs to category Y

1) Initialization, set initial values to each connection weight in the network

2) Select a new sample and it's desired output for training in random

3) Calculate the output value, error and global error of each hidden layer

4) If the global error is less than the threshold ε or reach the number of learning times K, the training is over, return (2)

5) Otherwise, use the weight update formula to adjust the weight, return (3)

6) When all sample training was completed, output the final result

Algorithm description:

DNN is a BP neural network that has several layers, and we adjust the connection weight of each layer in the process of model training continuously, which greatly improves the accuracy of forecasting results.

The corresponding fake code is as follows:

Input：X,Training samples
Input：M,The number of hidden layer
Input：N,Number of nodes
Input：ε,Error threshold Input：K,Number of iterations

---

initial
S = 1
Repeat
For $x \in X$ do

$$E = \frac{1}{2}\sum_{j \in A} e_j^2$$

$$\Delta\omega_{ji} = -\mu\frac{\partial E}{\partial \omega_{ji}} = \mu e_j \varphi_j'(X_j) y_i$$

$$\omega_{ji} = \omega_{ji} - \Delta\omega_{ji}$$

$$\Delta\omega_{ih} = -\mu\frac{\partial E}{\partial \omega_{ih}} = -\mu(\sum_{j \in A}\frac{\partial E}{\partial y_j}\varphi_j'(X_j)\omega_{ji})\varphi_i'(X_i) y_h$$

$$\omega_{ih} = \omega_{ih} - \Delta\omega_{ih}$$

End for
S = S + 1
Until E <= ε or s > N
Output：Model of DNN

V. EXPERIMENTS

A. Parameters Adjustment

In order to get the optimal result of the algorithm, we need to find the best parameters through experiments. Overall, there are some parameters that have large effects on the neural network algorithm result, such as the number of hidden layers, the number of hidden layer nodes, the learning rate and the number of learning times. While learning rate doesn't have an objective criteria, it needs us consider the actual data to set, so this paper uses adaptive learning rate. When the number of learning reaches a certain value, the result of model tends to be stable, thus the increase of the number of learning will only lead to training time and space wastage at this time. The number of model learning is set to 200 through simple experiment and we are going to determine the number of hidden layers and the number of hidden nodes in the next.

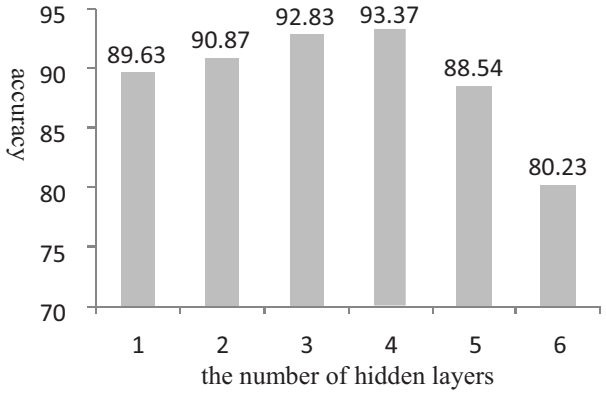

Fig. 2. The relationship between forecast accuracy and the number of hidden layers

It can be seen from Fig.2 that when the number of hidden layers is 4, the prediction accuracy of the model reaches the maximum value as 93%. In principle, when the number of hidden layers increases, the complexity of the model increases so that the classification of the data is more accurate. But when the hidden layer reaches a certain value (here is 4), continuing to increase the number of hidden layers will only cause overfitting, the accuracy of test set will decrease instead.

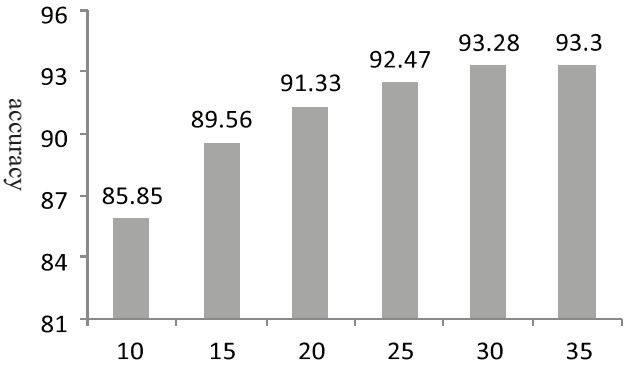

Fig. 3. The relationship between forecast accuracy and the number of hidden nodes

From Fig.3 we can see that when the number of hidden layer nodes reaches 30, the prediction accuracy of the model tends to be stable around 93%. Continuing to increase the number of hidden layer nodes at this point will lead to greatly reduction of the learning rate, while the improvement is not obvious. As a result, the number of hidden layers are determined to be 4, and the number of hidden layer nodes are determined to be 30.

B. Model Evaluation

Model evaluation is an important foundation for testing whether a model has enough value to use and it's an indispensable part of the model building process. In this paper, we will evaluate the model both on confusion matrix and ROC curve.

1) confusion matrix

Confusion matrix is used mainly for comparing the difference between model predictions and actual values, and it is a commonly used and effective model assessment method[12]. We use the dataset that contains 297 samples for model testing to compare the difference between the predictions and actual values.

TABLE I. CONFUSION MATRIX ABOUT PREDICTIVE VALUE AND ACTUAL VALUE

	Actual Malignant	Actual Benign
Predicted Malignant	**119**	**13**
Predicted Benign	**6**	**159**

As can be seen from Table.1, the prediction accuracy of the model reached 93%, where malignant nodules reached 90% and the recall reached 95%, which means that almost all patients with malignant nodules can be identified correctly. As a comparison, the hospital's real clinical data show that there are only 63% patients with malignant nodules actually in people who have diagnosed with malignant nodules and taken surgical treatment.

2) ROC curve

ROC (Receiver Operating Characteristic) curve is often used to find a balance between finding true positives and avoiding false positives, and the points on the curve represent the true positivity of different false positive thresholds. In simple terms, the larger the area under the ROC curve, the better the model performs[13].

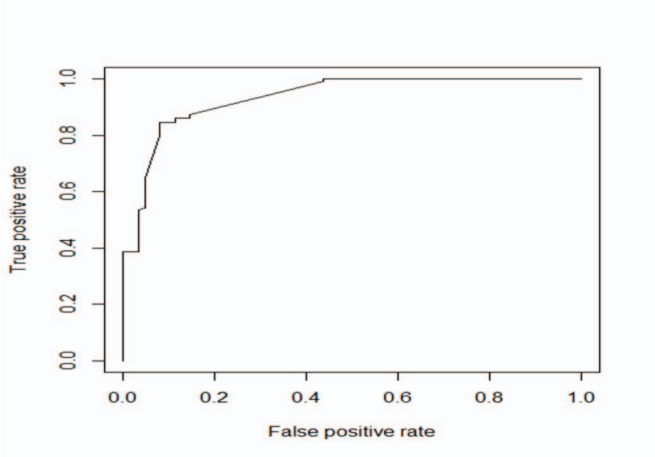

Fig. 4. ROC curve of Model

What we can see from fig.4 is that ROC curve is very close to the boundary of the rectangle, and the AUC (ROC curve area) is 0.935 while the maximum value is 1, which shows that our model has high performance.

C. Algorithm Comparison

In order to prove the superiority of our model, we choose several widely used and efficient data mining models to compare. In our experiments, the real clinical medical data and the new-thyroid dataset down from UCI were compared, and the result was as follows:

TABLE II. COMPARISON ON PREDICTIVE RESULTS AND REAL MEDICAL DATA SET WITH ALGORITHMS

	Precision	Recall	F value
RF	**88.22%**	**87.04%**	**0.88**
SVM	**87.42%**	**82.96%**	**0.85**
NN	**89.63%**	**77.04%**	**0.83**
DNN	**93%**	**96.64%**	**0.94**

TABLE III. COMPARISON ON UCI STANDARD DATA SET WITH ALGORITHMS

	Precision	Recall	F value
RF	**91%**	**87.84%**	**0.89**
SVM	**90%**	**91.89%**	**0.91**
NN	**92%**	**88.74%**	**0.90**
DNN	**95%**	**92%**	**0.93**

According to table 2 and table 3 above, we can see that DNN has a higher precision and recall on both dataset, which means that there is a kind of improvement by using DNN, and it is significant for diagnosis and prediction of diseases.

VI. CONCLUSIONS

Thyroid disease is a dangerous factor that threatening human health and the incidence rate has increased year by year, where thyroid nodules is one of the most harmful symptoms. How to improve the accuracy of diagnosis with benign and malignant nodules, and then recommend treatment measures correctly and timely for patients, has great significance both on patient's condition control and the savings of medical resources. A DNN algorithm is proposed in this paper shows a certain advantage both on the real medical dataset and the UCI standard dataset. It has higher accuracy and recall for forecasting, and the result is more convinced when taking the imbalance and sparseness of the actual medical dataset into account.

In the future, we would like to change the construction of DNN like constitutes of object function to adapt common medical data better, and improve the accuracy of diagnostic as much as possible.

ACKNOWLEDGMENTS

This research was supported by grants from Shanghai Science and Technology Innovation Action Plan (No.15511106900), Shanghai Science and Technology Development Funds (No.16JC1400802) and Shanghai Specific Fund Project for Informationization Development (XX-XXFZ-01-14-6349).

REFERENCES

[1] Ma H, King I, Lyu M R. Effective missing data prediction for collaborative filtering[C]//Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007: 39-46.

[2] Saif H, He Y, Alani H. Alleviating data sparsity for twitter sentiment analysis[C]. CEUR Workshop Proceedings (CEUR-WS. org), 2012.

[3] Masko D, Hensman P. The impact of imbalanced training data for convolutional neural networks[J]. 2015.

[4] Lusa L. Joint use of over-and under-sampling techniques and crossvalidation for the development and assessment of prediction models[J]. BMC bioinformatics, 2015, 16(1): 1.

[5] Ma J, Luo S, Dighe M, et al. Differential diagnosis of thyroid nodules with ultrasound elastography based on support vector machines[C]// Ultrasonics Symposium (IUS), 2010 IEEE. IEEE, 2010:1372-1375.

[6] Young Hun Lee, Dong Wook Kim, Hyun Sin In, Ji Sung Park, Sang Hyo Kim, Jae Wook Eom, Bomi Kim, Eun Joo Lee, Myung Ho Rho. Differentiation between Benign and Malignant Solid Thyroid Nodules Using an US Classification System[J]. Korean Journal of Radiology Official Journal of the Korean Radiological Society, 2010, 12(12):559-67.

[7] Kim E T, Park J S, Kim K G, et al. Computer-aided Diagnosis of Ultrasound Elastography for Classification of Benign and Malignant Thyroid Nodules[C]// Radiological Society of North America 2010 Scientific Assembly and Meeting. 2010.

[8] Rago T, Santini F, Scutari M, et al. Elastography: new developments in ultrasound for predicting malignancy in thyroid nodules[J]. The Journal of Clinical Endocrinology & Metabolism, 2007, 92(8): 2917-2922.

[9] Socher R, Karpathy A, Le Q V, et al. Grounded compositional semantics for finding and describing images with sentences[J]. Transactions of the Association for Computational Linguistics, 2014, 2: 207-218.

[10] Furuya S, Tominaga K, Miyazaki F, et al. Losing dexterity: patterns of impaired coordination of finger movements in musician's dystonia[J]. Scientific Reports, 2015, 5.

[11] Jin W, Li Z J, Wei L S, et al. The improvements of BP neural network learning algorithm[C]//Signal Processing Proceedings, 2000. WCCC-ICSP 2000. 5th International Conference on. IEEE, 2000, 3: 1647-1649.

[12] Deng X, Liu Q, Deng Y, et al. An improved method to construct basic probability assignment based on the confusion matrix for classification problem[J]. Information Sciences, 2016, s 340–341:250-261.

[13] Fawcett T. An introduction to ROC analysis[J]. Pattern Recognition Letters, 2006, 27(8):861-874.