IETE Zonal Seminar "Techno-Socio Development through Women Empowerment" – 2018
Special Issue of IJECSCSE, ISSN: 2277-9477

Survey on Involution of Neural Network

Sanjana R. Yamawar **Sapana D. Chaudhari** **Dipak S. Shirsode** **Prof. Dhiraj D. Shirbhate**

***ABSTRACT*-- The aging population of China is becoming increasingly more prominent, thus increasing the burden on medical resources. Therefore, the use of data mining technology to improve the efficiency of disease diagnosis has the following important significance. For hospitals, such technology can reduce the cost of providing one-on-one guidance to patients and the probability of registration errors. For patients, it can save time and energy spent on hospital visits in addition, through remote access patients can follow the automated guidance at home to complete registration, thereby enhancing admission efficiency. For internet users, such technology enables self-checking of these users' health conditions on a regular basis; based on certain main symptoms, possible diseases can be pre-diagnosed, thus providing a risk warning. Online medical guidance has become a very important step. To this end, we focus on employing the data mining technology to enhance the performance of online medical guidance. In this paper, we propose a medical diagnosis method called the named entity recognition method and a convolutional neural network model. We apply our proposed method and model as an innovative frame work for hospitalization guidance to provide human-like, comprehensive and informative automated medical consultations. We perform experiments on real-world datasets. The experimental results show that our methods achieve state-of-the-art performance compared with baselines.**

***Keywords*: Medical guidance, convolutional neural network, name entity recognition.**

I. INTRODUCTION

Nowadays, internet has become the key bridge for connecting patients/individuals with medical services. Whenever people do not feel well, nearby 90% of them first go and check symptoms on the internet to search for related medical information. The internet has changed the medical services for all major steps that may include medical consultation, clinic visits, treatment, as well as buying medication online.

According to a recent survey, internet healthcare will have a total market value of one trillion. Current medical systems at present already have self-diagnosis functions that may have registration systems available in certain medical systems. Though, the majority of medical network systems are totally based on expert systems, such as experts use their experience to pre-diagnose diseases that is based on patient's symptoms .Such systems require multiple experts, and consumes time, effort, and manpower.

This paper will focus on data mining and machine learning technologies to research medical guidance with the objective of providing human-like, automated medical consultation. Most individual when they become sick, as per their lack of medical knowledge and experience, they will frequently describe their symptoms inappropriate in medical terms.

The medical guidance model mentioned in this paper uses deep learning technologies, such as CNN and NLP, to perform transformation on raw and noisy data. The current model covers 500 different types of diseases.

II. LITERATURE REVIEW

The basic of medical guidance depends on machine diagnosis, which is a historic topic. As early as 1966, Ledley and Lusted proposed the idea of machine diagnosis. In 1972, Wilcox et al. also attempted to use a computer and applied Bayesian theory to identify a bacterial disease. In 2001,Saeys et al.proposed applying a feature selection technique to biological information, which was published in Bioinformatics (Oxford University Press). He proposed that, although the feature selection technique has wide application in the field of bioinformatics, in the biological field, this application of the technique has just been initiated. Because the medical samples exhibit the features of large dimension and short length (as in medical text information), it is necessary to modify and optimize the feature selection technique according to such characteristics of medical data.

In 2010, in an article also published in Bioinformatics, Abeel et al. studied a feature selection algorithm to identify biochemical features in the diagnosis of cancer and used a support vector machine classification algorithm to apply the integrated feature selection technique to disease diagnosis. In recent years, relevant studies on the text classification technique in the medical field have become progressively more mature. Concerning the relevant diagnosis of heart diseases and neural diseases, Ahmed integrated an artificial bee colony algorithm and a modified full Bayesian network classifier and used this combined technique for the mixed estimation, therein achieving a nearly 100% accuracy for heart disease prediction.

III. MEDICAL NAMED ENTITY RECOGNITION

In language usage, a named entity is defined as a text string with an independent meaning that is often used in a sentence. NER is the recognition of an entity having a specific meaning in the text, such as the named entity in the open field, that may include names of people, places, organizations and institutions. There may also be various classification in different sub-divided fields.

Mostly, the named entities in the medical field falls into four categories: disease, symptom, examination, and treatment. The artificial-intelligence disease diagnosis studied in this paper is mainly for recognizing named entities in the category of symptom.NER is a pattern recognition task, that means, identifying boundary information and type information of the entity from a given sentence. A typical method that can be used is to combine the boundary information and type information as a series of labels; then ,the task of NER is converted into forecasting a label for each word in the sentence. A typical labeling method generates labels in the form of B_C and I_C, where B and I are position labels, C is a category label, B is the beginning of an entity, and I is the continuation of an entity. Content that does not belong to any entity is generally labeled with an O.

The NER includes two classes of methods. One class that is based on classification, and the other class that is based on the serialization of annotations. Because methods based on the serialization of annotations are superior to the methods based on classification in many aspects, this paper adopts a method based on the serialization of annotations, i.e., the Conditional Random Field (CRF) model, to conduct the recognition of the symptom class among named entities in the medical field.

IV. CONVOLUTIONAL NEURAL NETWORK

The objective of Artificial-intelligence-guided disease diagnosis proposed in this paper is to extract the symptom description from the user inquiry and combine the symptoms to infer the most likely disease being suffered by patients under the combination of symptoms. In this report, we perform the mathematical abstraction of this question, Specifically, giving a series of symptoms to derive the disease classification for the patients exhibiting these symptoms.

In this paper, experts introduce the CNN model to solve this disease classification problem based on symptoms. CNN always compare image piece by piece. Whenever it is presented with a new image, the CNN doesn't know what and where exactly these feature will match so it tries it everywhere in every position. First we annotate the extracted disease described by users with a natural language disease inquiry through named entity recognition (NER).Then we use word embedding to convert the disease inquiry data from users for the matrix expression.

V. EXPERIMENTS

Here we are going to see that, experts evaluate the effectiveness of the proposed approach for online medical guidance with experiments on huge real-world data sets. All the data used in this paper are public data from the internet and have been acquired from web. In particular, the data involving questions/answers to/from doctors and the disease and symptom data are obtained from several major medical information from websites. A detailed description of the data is provided in Table 1.

Dataset	Introduction	Data Amount
1	Disease Information 1	7835
2	Disease Information 2	489
3	Symptom Information	6609
4	Question Answer 1	3263714
5	Question Answer 2	81965

Table 1:Description of Medical Data

For the disease information data set and the symptom information dataset, experts always extract the disease names and symptom names from the datasets and take the correlation between diseases and symptoms as the knowledge base, which may play a significant role in the diagnosis model. For the question-answer dataset, experts extract the content, including the titles of user inquiry, user properties, content of user inquiry, and they also take answers from doctors. However, because the relevant data in the medical field are professional property, some questions from patients obviously include network language, and the word segmentation results for some words are not satisfactory. Therefore, in this paper, experts use the method of new word extraction based on mutual information and extract portions of new words that cannot be recognized by the tools for word segmentation.

1. WORD EMBEDDING TRAINING

The meaning of ''word embedding'' is to embed a word into a vector space, namely, using a vector to express the word. Again in contrast to the Vector Space Model (VSM), word embedding also uses many vector expressions for the training words without term supervision to make this vector expression rich in semantic information. Therefore, the word is expressed as a vector with a relatively low dimension; meanwhile, this vector also has certain abstract semantics.

Phrase(in Chinese)	Phrase (in English)	Similarity
Laduzi	Diarrhea	0.796444
Laxi	Diarrhoea	0.747393
Futong	Bellyache	0.654137
Fuxie	Diarrhoea	0.709832
Bianmi	Astriction	0.635700
Outu	Anabole	0.588085

Table: Sample of similarity using word embedding.

A vector's abstract semantics can be expressed such that, for two words with very similar meaning, even the similarity of their characters is very small; for example, such that the vector expression of ''diarrhea'' and ''enterorrhea'' is very close, and the similarity is very high in the vector space. In addition, word embedding can also be expressed as the relationship between words; one classic example that can be taken is the relationship between the four roles of queen, king, man, and woman: V queen−V women +V man ≈V king.

In this paper, experts will apply the word embedding training tool from the Skip-Gram model , to train the word embedding on the medical question/ answer text. For the obtained word embedding, can use the cosine similarity to compute the degree of similarity. Table lists the Top similar words to the Chinese word ''Fuxie'', which means ''diarrhea''.

2. STATISTICAL ANALYSIS OF THE DATA

For the 7,835 pieces of disease data in dataset 1, experts use the disease names to match the data after combining dataset 4 and dataset 5. If there is a matched disease, experts correlate this inquiry with the disease. The top matched diseases and their number of matches are listed in Table. The data statistics reveal that the diseases that rank at the top of the statistics are all common diseases. According to the statistics, the cumulative percentage of disease data whose number of matches ranks in the first 500 is 92.5%.

Disease	Amount	Precision %
Common Cold	99483	2.973477133
Menoxenia	99170	2.964121782
Missed Miscarriage	92325	2.75952953
Miscarriage	74799	2.235689676
Vaginitis	67358	2.013283402
Vitiligio	51147	1.5284747976

Table : for disease match.

3. NAMED ENTITY RECOGNITION

Before recognizing the named entity of the medical information, experts should extract the features of the named entity. The design of the feature extraction template depends on the format of the input file. In particular, there are two types of features:

Unigram features and Bigram features. For the Unigram features each row %x[#,#] means to generate a function of a point in the CRF, f(s,o), where s is the label at time t and o is the context at time t. For the Bigram feature, each row %x[#,#] means to generate a function of the side in the CRF, f(s0,s,o), where s0 is the label at time (t −1).

The Bigram feature is often simplified to a B in the template, and then, the default generation f(s0,s)(namely ,the previous output label and the current output label) is combined as the Bigram feature. Table 2 shows an example of the input file format used in the symptom recognition process,

VI. ADVANTAGES AND DISADVANTAGES

Advantages

- It require less formal statistical training.
- Ability to implicitly detect complex nonlinear relationships between dependent and independent variables.
- Ability to detect all possible interactions between predictor variables.
- Availability of multiple training algorithms.

Disadvantages

- Greater computational burden. Proneness to greater over fitting.
- Empirical nature of model development.

VII. CONCLUSION

In this paper, we have introduced a novel framework of medical guidance. Specifically, that experts first annotated the extracted disease described by the users with the natural-language disease inquiry through named entity recognition (NER). Then, we have encountered word embedding to convert the disease inquiry data of users for the matrix expression. Finally, we have proposed a model for artificial-intelligence-guided disease diagnosis and

eventually provided intelligent guidance for disease inquiry for individuals. We have evaluated our methods with extensive experiments on real-world datasets. The experimental results clearly validate the effectiveness of our methods. Potentially, this study has many future research directions. First, it would be interesting to investigate new guide diagnosis models to improve the accuracy of online medical guidance.

VIII. REFERENCES

1. S.Mertens,F.Gailly,andG.Poels,''Supporting and assisting the execution of flexible healthcare processes,'' in Proc. Int. Conf. Pervas. Comput. Technol. Healthcare, 2015, pp. 375–388.
2. Y. Zhang, L. Sun, H. Song, and X. Cao, ''Ubiquitous WSN for healthcare: Recent advances and future prospects,'' IEEE Internet Things J., vol. 1, no. 4, pp. 311–318, Aug. 2014.
3. M. R. Friesen and R. D. McLeod, ''A survey of agent-based modeling of hospital environments,'' IEEE Access, vol. 2, pp. 227–233, 2014.
4. A.T.Sadiq and N.T.Mahmood, ''Ahybrid estimation system for medical diagnosis using modified full Bayesian classifier and artificial bee colony,'' Iraqi J. Sci., vol. 55, no. 3A, pp. 1095–1107, 2014.
5. R. R. Patil, ''Heart disease prediction system using naive Bayes and Jelinek-mercer smoothing,'' Int. J. Adv. Res. Comput. Commun. Eng., vol. 3, no. 5, pp. 6787–6789, 2014.
6. A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks,'' in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

AUTHOR'S PROFILE

Sanjana R. Yamawar

BE CSE J.D.I.E.T. YAVATMAL
Research work 1 on

SMART SPECTS FOR VISUALLY IMPAIRED PEOPLE

Sapana D. Chaudhari

BE CSE J.D.I.E.T. YAVATMAL
Research work 1 on

SMART SPECTS FOR VISUALLY IMPAIRED PEOPLE

Dipak S. Shirsode

BE CSE J.D.I.E.T. YAVATMAL
Research work 1 on

DATA MINING TECHNIQUES FOR FRAUD DETECTION

Asst. Prof. Dhiraj D. Shirbhate

Asst. Prof. J.D.I.E.T YAVATMAL
ME CSE ,Publications 10 on Database system ,

Research work 4 paper.