Study of Data Representation Methods for TCM Clinical Assistant Diagnosis

Yin Liu¹*, Hongguang Chen¹

¹Department of Software Engineering, Chengdu University of Information Technology, Chengdu, Sichuan, 610225, China

*Corresponding author’s e-mail: liuyincn@foxmail.com

Abstract. The unbalanced distribution of medical resources renders the research on TCM (Traditional Chinese Medicine) clinical assistant diagnosis more important to regions with less medical resources. In recent years, more and more clinical assistant diagnosis methods are deep learning (DL) based. The input data representation of these DL models is one of the most important factors for achieving better results. In this paper, different data representations methods are investigated using a multi-layer perceptron for a multi-class multi-label TCM clinical assistant diagnosis task. From the experimental results, it can be concluded that fast-Text representation is more suitable to this task since TCM clinical records are brief with limited information.

1. Introduction

At present, clinical diagnosis of traditional Chinese medicine (TCM) in China relies heavily on the theoretical knowledge and clinical experience of doctors. For small hospitals with insufficient medical experts and resources, the misdiagnosis rate of difficult diseases is much higher than that of major hospitals. In order to obtain more accurate diagnosis, researchers have paid more attention to the task of TCM clinical assistant diagnosis.

Clinical assistant diagnosis of TCM is based on clinical records. However, the records are of different styles and lengths, which can lead to ambiguity in their interpretation. Data representation can be used to help with such problems. Data representation refers to the mapping process of original data. Since it’s impossible to make computers understand plain text naturally, the original data needs to be mapped into a certain feature space and converted into numerical or vector forms. After mapping, their similarities can be obtained by calculating their distances, etc. This solves the problems of semantic ambiguity for TCM clinical records, hence data representation has become the basis for TCM clinical assistant diagnosis research.

Data representation are divided into vector space representation and distributed representation [1]. The vector space representation are easy to interpret but limited by vocabulary size. The dimensions of the word vector goes exponentially higher as vocabulary size grows bigger, which will increase training time. However, since the represented data is sparse, the characteristics of the weights are prominent, so models may perform well with fewer iterations. On the other hand, distributed representation has dense numerical values, which makes it rich in semantic information, and its lower dimensionality makes it less resource-extensive. Hence distributed representation is widely used in deep learning-based data representation. However, it’s difficult to interpret, and achieving rich semantic information requires much longer time. Therefore, it is necessary to evaluate and choose a suitable data representation method for TCM clinical assistant diagnosis.
For evaluation, we first process the data using different selected data representation methods, and then compare their performances in other primary tasks. These tasks can be text classification [2], part-of-speech tagging [3], text generation [4] and information retrieval [5]. The TCM clinical records can be regarded as natural labels, which makes TCM clinical assistant diagnosis a multi-label classification task.

2. Related Work
Most deep learning model cannot process plain text directly. The input data need to be converted into numerical forms before training. Therefore, data representation of original data has become the basis of deep learning. This section introduces two methods of data representation: vector space representation and distributed representation.

In early NLP tasks, vector space representation was used to represent word vectors, and one-hot encoding is used primarily [6]. One hot encoding is a process of encoding categorical features as a one-hot numeric array. It’s concise and fast, but 0s and 1s are unable to capture the importance of individual word, hence some propose that the weights are replaced with word frequency [6]. Other researchers have used tf-idf, mutual information value (MI), etc. to represent word weights [7]. However, the vectors can become very large. In addition, the one-hot encoded forms of synonyms and near-synonyms are very different, which may lead to poor accuracy of semantic representation.

To solve these problems, researchers have studied distributed representation for word vectors, which was first proposed by Hinton [8] in 1986. Word vectors with distributed representation have lower dimensions and are more semantically accurate. In addition, word vectors obtained by neural networks can perform linear transformations, such as V ("Queen")-V ("Woman") approximating to V ("King")-V ("man"). This plays an important role in applications such as semantic network, semantic disambiguation and relationship conversion. One such representative method is word2vec [9], which is a group of related models that are used to produce word embeddings. It can utilize either of two model architectures: continuous bag-of-words (CBOW) or continuous skip-gram. In CBOW the model predicts the current word \( w \) from a window of surrounding context words \( context (w) \), while the skip-gram model uses the current word \( w \) to predict context \( (w) \). Facebook proposed fast-Text, a state of the art distributed representation method [10]. Similar to CBOW, it uses n-gram information as input, which makes the word vector richer in information and faster to train. However, distributed representation requires large amount of data. Insufficient data will cause the quality of the word vector model to deteriorate.

3. TCM Clinical Assistant Diagnosis based on Deep Learning
Under the framework of deep learning (DL), DL-based TCM clinical assistant diagnosis methods can be divided into steps of data representation and model building.

3.1 Data Representation
Data representation transforms the original data into numerical values, usually in the form of multidimensional vectors. In TCM clinical assistant diagnosis, when the TCM clinical record is processed, a fixed-size vocabulary \( D \) is generated. Each word \( i \) in the vocabulary \( D \) will be mapped into a multi-dimensional vector to form the word vector \( v_i \). This will result in a word vector table \( R^{d \times |D|} \), where \( v \in R^{d \times |D|} \), \( D \) is the word vector dimension, and \( |D| \) is the vocabulary size.

3.1.1 Vector Space Representation. In vector space representation, assuming that \( i \) is the index value of the word \( i \) in the vocabulary \( D \), each dimension \( d_j \) in \( v_i \) has a weight \( w_{ij} \), where \( d_j \) represents the \( j \) dimension of the word vector \( v_i \), \( j \in \{1, 2, \ldots, |D|\} \). For weight value \( w_{ij} \), this paper selected one-hot representation, word frequency representation and mutual information value representation for evaluation.

One-hot representation: each word has only two possibilities in the word vector, occurrence and non-occurrence, so 0s and 1s are used to represent the two state. 0 indicates occurrences and 1 indicates non-occurrences. It can be expressed as follows:
\[ w_{ij} = \begin{cases} 1 & i = j \\ 0 & \text{otherwise} \end{cases} \] (1)

Word frequency representation: different from one-hot representation, it replaces 1s with word frequency \( t_{fi,j} \), which can be expressed as:

\[ w_{ij} = \left\{ \begin{array}{ll} t_{fi} & i = j \\ 0 & \text{otherwise} \end{array} \right., \text{where } t_{fi} = \frac{n_i}{N} \] (2)

Here, \( N \) is the total number of words in the training data set, and \( n_i \) represents the number of words \( t_i \) in the training data set. Word frequency indicates the probability of words appearing in the entire data set. Note that sometimes words with lower frequencies are also very important.

Mutual Information Representation (MI): improving upon word frequency representation and considering the relationship between the word and label \( c \) \((c \in C, C \) represents all the label set\), is calculated with respect to average mutual information \( MI_{ijc} \), which is expressed as follows:

\[ l_{ic} = t_{fi} * \log \frac{t_{fi} * t_{ic}}{t_{fi} * t_{c}} \] (3)

\[ w_{ij} = \left\{ \begin{array}{ll} MI_{ijc} & i = j \\ 0 & \text{otherwise} \end{array} \right., \text{where } MI_{i} = \frac{1}{|C|} \sum_{c \in C} l_{ic} \] (4)

Here, \( t_{fi,c} \) represents the probability that the label \( c \) and the word \( t_i \) appear simultaneously, \( t_{fi,c} \) represents the probability that label \( c \) appears. \( l_{ic} \) represents the mutual information between label \( c \) and word \( t_i \), and \(|C|\) represents the number of label types. Mutual information indicates the relationship between words and labels, but class-label information doesn’t necessarily reveal word information.

3.1.2 Distributed Representation. Distributed representation utilizes neural networks to represent words as lower-dimensional, densely-valued floating-point vectors. It captures the grammatical and semantic relationships of the words very well. In this paper we selected the skip-gram model of word2vec [9] and fast-Text [10] for evaluation.

In the continuous skip-gram architecture, the model uses the current word to predict the surrounding window of context words, where each word is represented as a bag of character n-grams. In fast-Text, however, a vector representation is associated to each character n-gram, and the words are represented as the sum of these representations.

The numerical denseness of the distributed representation means the model can contain more semantic information, but it also suffers from poor interpretability. However, it is faster with smaller number of parameters, making it more time efficient.

3.2 Multi-label Classification based on Multi-layer Perceptron
TCM clinical records can be regarded as natural labels, which makes TCM clinical assistant diagnosis a multi-label classification task. In this paper, a multi-layer perceptron (MLP) based on a simple multi-class classification neural network was used [11]. It consists of one input layer, three hidden layers, and one output layer. The input of the model is a vector representation of the TCM clinical records using the word vector model. The output is a multi-dimensional vector whose size is the number of disease labels. The activation function for the output layer is softmax, therefore each dimension of output vector can be used to represent the probability of predicted labels. The performance of different data representation methods is evaluated based on these output results.

4. Experiment

4.1 Experimental Data and Experimental Settings
A total of 10,000 authentic TCM clinical consultation records randomly selected from a medical information system were used in the experiment. The dataset consists of 7,500 training data with 996 labels and 2,500 test data with 4,695 labels. The vocabulary size is 7,138.
4.2 Evaluation Method
The PRF evaluation criteria was used in this paper: Precision \( (P) \), Recall \( (R) \), and F-score \( (F) \), shown below:

\[
P = \frac{|NBPR|}{|NP|}, \quad R = \frac{|NBPR|}{|NBP|}, \quad F = \frac{2PR}{P+R}
\]

Here, \(|NBPR|\) represents the number of correctly classified label in the prediction result, \(|NP|\) represents the number of label in the prediction result, and \(|NBP|\) represents the number of label in the actual class. The F-score evaluates the overall performance of the experiment.

4.3 Experimental Results and Analysis
In the experiment, two regularization methods (L1, L2) and three update strategies (SGM, NAG and Adam) were selected for evaluation. The record with less than 4 labels accounts for 98.75% of the entire data set, therefore we choose the PRF score of TOP-3 prediction result for comparison. The results are shown in Table 1.

| Table 1. F-score of five data representation methods. L1, L2, SGM, NAG and Adam are used as control variables. |
|---------------------------------------------------------------|
|                  | L1          | L2          | SGM         | NAG         | Adam         |
| one-hot           | 0.39057     | 0.40036     | 0.41297     | 0.40766     | 0.45561     |
| Frequency          | 0.17570     | 0.17371     | 0.19164     | 0.18914     | 0.21951     |
| MI                | 0.21022     | 0.11183     | 0.11183     | 0.21022     | 0.21934     |
| skip-gram Model    | 0.11763     | 0.11415     | 0.18450     | 0.18367     | 0.19761     |
| fast-Text          | 0.13555     | 0.18616     | 0.28405     | 0.28438     | 0.29085     |

According to Table 1, for vector space representation, one-hot encoding achieves the best results. Since it produces weights of equal probability, it’s more effective when used on short text that has uniformly distributed features such as TCM clinical records. For distributed representation, fast-Text obtains better result, because of its textual n-gram representations. In the clinical records of TCM, different combinations of words may lead to misdiagnosis, with n-gram semantics of features, the represented information is more clear, so fast-Text representation fares better than skip-gram.

Time complexity should also be taken into consideration. According to the training time statistics, models using distributed representation methods takes an average of 932s, compared to the 15,544s by vector space representations, which is approximately a 17-fold difference. This is because vector space representation has too many parameters due to high dimensionality, therefore it’s more time consuming and resource extensive.

Figure 1. (a) loss of one-hot using Adam, (b) loss of fast-Text using Adam. The horizontal axis represents the number of iterations, and the vertical axis represents the loss value.

When the models are completely converged, the F-score for fast-Text is 0.42226, which is increased by 45.18%; one-hot is slightly decreased at 0.44698. The results are shown in Figure 1. Fast-Text and one-hot achieved comparable results here. However, the rich semantic information in fast-Text makes the represented information more clear and meaningful for the task. Considering the time complexity,
when dealing with short text with limited information such as TCM clinical records, fast-Text representation is more suitable for this task.

5. Conclusions
This paper studies the performance of different data representation methods for the task of TCM clinical assistant diagnosis. These representation methods are evaluated through a multi-label classification task. Since TCM clinical records are short and limited in information, one-hot encoding from vector space representation can perform well on this task. However, fast-Text from distributed representation can achieve comparable results with much less training time. Data representation is the basis of deep learning-based natural language processing tasks. Future work needs to focus on combining TCM clinical records and distributed representation methods to improve the experimental results for further research.

References
[1] Smith N. A. Contextual Word Representations: A Contextual Introduction[J]. 2019.
[2] Ge L. Improving Text Classification with Word Embedding[C]// IEEE International Conference on Big Data. IEEE, 2018.
[3] Marulli F., Pota M., Esposito M. A Comparison of Character and Word Embeddings in Bidirectional LSTMs for POS Tagging in Italian[J]. 2018.
[4] Arslan Y., Dilek Küçük, Birturk A. Twitter Sentiment Analysis Experiments Using Word Embeddings on Datasets of Various Scales[C]// International Conference on Applications of Natural Language to Information Systems (NLDB). Springer, Cham, 2018.
[5] Manning, Christopher D., Raghavan, Prabhakar, Schutze, Hinrich. Introduction to Information Retrieval || Evaluation in information retrieval[J]. 2008.
[6] Turney P. D., Pantel P. From Frequency to Meaning: Vector Space Models of Semantics[J]. Journal of Artificial Intelligence Research, 2010, 37(1):141-188.
[7] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.
[8] Hinton G. E. Learning distributed representations of concepts.[C]// Eighth Conference of the Cognitive Science Society. 1989.
[9] Mikolov T., Chen K., Corrado G., et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.
[10] Bojanowski P., Grave E., Joulin A., et al. Enriching Word Vectors with Subword Information[J]. Transactions of the Association for Computational Linguistics, 2017, 5:135-146.
[11] Hannan M. A., Arebey M., Begum R. A., et al. An automated solid waste bin level detection system using Gabor wavelet filters and multi-layer perception[J]. Resources, Conservation and Recycling, 2013, 72:33-42