Traditional Chinese Medicine Clinical Records Classification using Knowledge-Powered Document Embedding

Liang Yao, Yin Zhang*, Baogang Wei, Zherong Li and Xiangzhou Huang
College of Computer Science and Technology
Zhejiang University, 310027 Hangzhou, China
Email: {yaoliang, yinzh, wbg, rsmile_lzr}@zju.edu.cn, 690003374@qq.com
*Corresponding Author

***Abstract*—Text classification is one of the fundamental tasks in text mining. In the medical domain, there have been a number of studies on text classification in modern medicine clinical notes written in English. However, very limited text classification research has been conducted on clinical notes written in Chinese, especially traditional Chinese medicine (TCM) clinical records. The goal of this study was to investigate features and machine learning classification algorithms for TCM clinical text classification. We collected 7,037 TCM clinical records of famous TCM doctors as our dataset, and investigated the effects of different types of features and classification algorithms. Additionally, we proposed a novel method to combine deep learning text representation with TCM domain knowledge, which results in the best classification performance.**

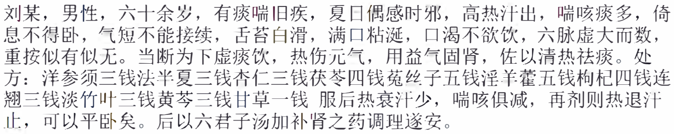
***Keywords*—Traditional Chinese Medicine, Clinical Records Classification, Features, Classification Algorithms, TCM domain knowledge**

I. Introduction

As a complete medical knowledge system other than orthodox medicine, traditional Chinese medicine (TCM) plays an indispensable role in the health care for Chinese people for several thousand years [7]. In TCM, huge amount of clinical records in the ancient textbooks or hospitals are the main TCM knowledge sources for the generation of appropriate clinical hypotheses. Fig. 1 shows a sample of clinical records, it describes a patient's chief complaint and history, how a doctor diagnoses and prescribes a prescription and its effects.

As a fundamental task of clinical text mining, clinical text classification plays an important role in clinical records organization, retrieval and so on. In recent years, many studies on clinical text classification focus on modern medicine clinical text written in English. However, studies focusing on clinical text in Chinese are relatively limited, especially on TCM clinical records.

In this paper, we investigate features and classification algorithms for TCM clinical text classification. Using 7,037 TCM clinical records of famous TCM doctors, we evaluated the contribution of different types of features and classification algorithms for text classification in TCM clinical text. In addition, we propose a novel feature generation method which combines a deep learning text representation algorithm *doc2vec* [4] and TCM domain knowledge. The generated features are effective for TCM clinical text classification.



刘某，男性，六十余岁，有痰喘旧疾，夏日偶感时邪，高热汗出，喘咳痰多，倚息不得卧，气短不能接续，舌苔白滑，满口粘涎，口渴不欲饮，六脉虚大而数，重按似有似无。当断为下虚痰饮，热伤元气，用益气固肾，佐以清热祛痰。处方：洋参须三钱法半夏三钱杏仁三钱茯苓四钱菟丝子五钱淫羊藿五钱枸杞四钱连翘三钱淡竹叶三钱黄芩三钱甘草一钱 服后热衰汗少，喘咳俱减，再剂则热退汗止，可以平卧矣。后以六君子汤加补肾之药调理遂安。

Fig. 1: An example TCM clinical record.

The main contributions of the study are summarized as: (1) To the best of our knowledge, this is one of the earliest comprehensive studies on features and classification algorithms for TCM clinical text. (2) We propose to combine deep learning text representation methods with TCM domain knowledge, which leads to the best results. (3) To our knowledge, this is one of the earliest works to apply knowledge-based deep learning techniques for text classification, which may inspire some further research of general text mining. (4) We provide a benchmark TCM clinical record dataset[1].

II. Method

A. Data

We collected 7,037 TCM clinical records from *Classified Medical Records of Distinguished Physicians Continued Two* (*Er Xu Ming Yi Lei An* in Chinese, ISBN 7-5381-2372-5). The 7,037 records are in 5 categories (Internal Medicine, Surgery, Gynaecology, Ear-Nose-Throat & Stomatology and Paediatrics). We randomly split the dataset into training set and test set. The training dataset contains 4,882 records and the test dataset contains 2,155 records. We show the number of records in each category in Table I.

B. Features

1) Bag of words (BOW): The simplest representation of documents. We segment clinical records into words by using a popular Chinese natural language processing tool FudanNLP[2] and a TCM vocabulary contains about 180k words collected by us. After word segmentation, we remove words appearing less than 5 times, which results in 11,954 words left. We

[1] We released the dataset and source code of this paper at https://github.com/yao8839836/CEMRClass.

[2] https://github.com/xpqiu/fnlp/

978-1-5090-1610-5/16/$31.00 ©2016 IEEE

TABLE I: Statistics of the dataset.

Category	# of clinical records	
	Training	Test
Internal Medicine	1905	896
Surgery	574	233
Gynaecology	1044	423
Ear-Nose-Throat & Stomatology	522	241
Paediatrics	837	362

use binary vectors, term frequency (TF) vectors and term frequency-inverse document frequency (TF-IDF) vectors. We found different number of rare words don't change the results much.

2) Bag of concepts (BOC): A knowledge-based feature. We employ two TCM knowledge sources: Traditional Chinese Medical Subject Headings (TCM MeSH) [8][3] which is compatible with Medical Subject Headings (MeSH) and Clinic terminology of traditional Chinese medical diagnosis and treatment (GB/T 16751-1997)[4] which is a national standard of China. Most concepts in the two knowledge sources have some descriptive words. We use 8,307 concepts with different types from TCM MeSH, and 930 diseases, 487 syndromes from GB/T 16751-1997. We represented clinical records as binary vectors indexed by concepts. We identify concepts by simply using text matching.

3) Explicit semantic analysis (ESA): A knowledge-based feature. The original ESA method represents a word as the vector of all Wikipedia articles, each position of the vector is the TFIDF value of the word in a Wikipedia article [3]. Similarly, we use 9,649 concepts which have descriptions in 8,307 TCM MeSH concepts and 1,417 GB/T 16751-1997 concepts, each word in clinical records is represented as a 9,649 dimensions vector, and each position is the TFIDF value of the word in the 9,649 documents description corpus. Formally, the ESA vector of a word w is defined as:

$$ESA_w = \{TFIDF_{c_1}, TFIDF_{c_2}, \dots, TFIDF_{c_M}\} \quad (1)$$

where each $TFIDF_{c_m}$ is the TF-IDF value of w in concept description document of c_m, TF is the term frequency in description document of c_m, IDF is inverse document frequency in the description corpus of all employed concepts. We then use the average ESA vector of all words appearing in the description corpus as the feature vector of a record.

4) LDA features: Latent Dirichlet Allocation (LDA) [1] is a widely used topic model, which represents each document as a probability over a set of "topics" characterized as different distributions over vocabularies. LDA-based approaches achieve good performance in terms of capturing the semantics of texts in several classification tasks. We use the topic distribution of each clinical record as the feature vector. After word segmentation, we remove stop words and words that appear less than 10 times, and get a vocabulary of 6,699 words. We set the number of topics $K = 100$, hyperparameters $\alpha = 50/K$, $\beta = 0.1$. We found different number of rare words and parameters don't change the results much.

5) word2vec features: A number of studies have been devoted to building distributed word embeddings which can encode both syntactic and semantic information of words into continuous vectors. Recently, the well-known *word2vec* model [5] (including continuous bag-of-word (CBOW) model and the continuous Skip-Gram model) has been proposed. The training objective of CBOW is to combine the embeddings of surrounding words as input to predict the central word in a sliding window; while Skip-Gram tries to use the central word as input to predict the surrounding words in a sliding window. We train Skip-Gram and CBOW on all clinical records to obtain word vectors, then use the TF-IDF weighted average of all words' vector in a clinical record as the feature vector.

6) doc2vec features: doc2vec models are document embedding models proposed recently [4], including the distributed memory model (PV-DM) based on CBOW and the distributed bag of words model (PV-DBOW) based on Skip-Gram. *doc2vec* models first train CBOW or Skip-Gram in *word2vec* to obtain word vectors, then treat a document id as a word in every context window and train the similar *word2vec* model again to obtain the document vector, but not update pre-trained word vectors. *doc2vec* models are reported to achieve the state-of-the-art performance on document classification. We use the same setting as word2vec features.

7) The proposed knowledge-based word2vec features: We combine *word2vec* model with ESA knowledge obtained from TCM MeSH and GB/T 16751-1997. The basic idea is that *word2vec* model trained on clinical records only considers the context information of words, but ignores TCM domain knowledge, while ESA vectors of word encode domain knowledge, and could help *word2vec* training process. Similar words in TCM knowledge should also be similar in the word vector space. We name our Skip-Gram based model ESA-Skip-Gram and CBOW based model ESA-CBOW.

Specifically, we compute the cosine similarity of every word pair (w_x, w_s) using ESA vectors as following:

$$r(w_x, w_s) = \frac{ESA_{w_x} ESA_{w_s}}{|ESA_{w_x}||ESA_{w_s}|} \quad (2)$$

We then use the top k similar words of w_x as the must-link set S_{w_x} of w_x. In our experiment, we set $k = 100$. As an example, we can add regularization to the original Skip-Gram objective. Given a word sequence $w_1, w_2, \dots, w_X$, the objective of the ESA-Skip-Gram is to maximize:

$$\begin{aligned} L = \frac{1}{X} \sum_{x=1}^{X} \sum_{-N \le c \le N, c \ne 0} \Big(\log p(w_{x+c} | w_x) \\ + \sum_{w_s \in S_{w_x}} r(w_x, w_s) sim(\mathbf{v}_{w_x}, \mathbf{v}_{w_s}) \Big) \end{aligned} \quad (3)$$

where the first term in the bracket is the Skip-Gram objective, the second term is the regularization. w_x is the central word, w_{x+c} is a surrounding word, and N indicates the context window size to be $2N + 1$, w_s is a similar word of w_x, $sim(\mathbf{v}_{w_x}, \mathbf{v}_{w_s})$ is the cosine similarity of word vector $\mathbf{v}_{w_x}$ and word vector $\mathbf{v}_{w_s}$. Updates for $\mathbf{v}_{w_{x+c}}$ and $\mathbf{v}_{w_x}$ are the same as Skip-Gram, the difference is after updating $\mathbf{v}_{w_{x+c}}$ and $\mathbf{v}_{w_x}$, we update vector of each word w_s in S_{w_x} as: $\mathbf{v}_{w_s} \leftarrow \mathbf{v}_{w_s} - \alpha_s \sum_{-N \le c \le N, c \ne 0} r(w_x, w_s) \cdot \Big(\frac{\exp\{f(w_{\boldsymbol{p}})\}}{\exp\{f(w_{\boldsymbol{p}})\}+1} -$

[3] http://zcy.ckcest.cn/tcm/dic/home

[4] http://zcy.ckcest.cn/tcm/dic/sr

$I[w_{x+c} = w_p]\Big) \cdot \mathbf{v}_{w_{x+c}}$, where w_p is the central word predicted by the model, $I[x]$ is 1 when x is true, $f(w_p) = \mathbf{v}_{w_{x+c}} \cdot \mathbf{v}_{w_x}$, α_s is the learning rate.

8) The proposed knowledge-based doc2vec features: We first train knowledge-based *word2vec* in the last sub-section, then use the same method to generate document (clinical record) vectors using *doc2vec* models. We name our PV-DBOW based model ESA-PV-DBOW and PV-DM based model ESA-PV-DM.

C. Classification Algorithms

In this study, we compared six state-of-the-art classification algorithms: non-linear kernel SVM (LibSVM), linear kernel SVM (LibLinear), maximum entropy (MaxEnt), naive Bayes (NB), C4.5 and random forest (RF) which are commonly used in many classification tasks. These classifiers are the most competitive ones for real world classification problems [2].

III. Results

We evaluate the performance of the clinical records classification by using macro-averaged F_1 score [6]. Table II shows the classification results of each feature space under each classifier. We can observe that the most competitive features are our ESA-PV-DBOW features and BOW features, and the highest macro-averaged F_1 score 0.8176 was achieved by ESA-PV-DBOW and LibSVM. BOW is simple, and it achieves high macro-averaged F_1 score, but the dimension is very high, which will lead to high time and memory cost, while ESA-PV-DBOW takes the efficient hierarchical softmax in *word2vec*, and the dimension is much lower. We can also note that PV-DBOW and Skip-Gram perform better than CBOW and PV-DM, which means Skip-Gram based methods capture more semantics of clinical records. After adding TCM knowledge, all four embedding methods perform better, which shows the effectiveness of TCM knowledge encoded by ESA. BOC and ESA perform steady and achieve some relatively good results, which means it also makes sense when using TCM knowledge only. LDA's performance is not very satisfactory, clinical records have many topical words in common, which makes them share similar topic distributions, and weakens distinctive capabilities of topics. Table II also presents performances of different classifiers. From the table, we can see that LibSVM, LibLinear and MaxEnt have better predictive capabilities than others in general, and LibSVM is the best classifier for high quality features like ESA-PV-DBOW and BOW.

Table III gives F_1 scores for each category using several high quality features and LibSVM. We can note that gynaecology records are the most distinctive ones, for they have many contents only for female patients. Paediatrics records get the worst performance, for children and babies share many diseases and symptoms with adults, the classification might be misled. Categories which have fewer records are more difficult to be classified. We could improve classification performance by oversampling records in rare categories.

IV. Conclusion and Future Work

In this study, we investigated features and classification algorithms for TCM clinical text classification. We further propose a knowledge-based deep learning method for clinical records representation, and achieve the best performance, which indicates a promising start for TCM clinical text classification research. The performance could be improved further by concatenating features (e.g., ESA-PV-DBOW feature and BOC feature) or adding other types of knowledge (e.g., hierarchical structure in TCM knowledge base), we plan to investigate these in our future work.

TABLE II: Macro-averaged F_1 scores of all types of features.

Feature	Dimension	Classifier					
		LibSVM	LibLinear	MaxEnt	NB	C4.5	RF
BOW Binary	11954	0.7391	0.7237	0.7479	**0.5870**	**0.5843**	0.4454
BOW TF	11954	0.7127	0.7220	0.7012	0.4496	0.5837	0.4612
BOW TFIDF	11954	0.7552	**0.7511**	0.7012	0.4505	0.5837	0.4529
LDA	100	0.4886	0.6076	0.6494	0.4943	0.4779	0.5906
ESA	9649	0.6543	0.7098	0.6750	0.5200	0.4649	0.3757
Skip-Gram	200	0.6079	0.6490	0.7101	0.4479	0.4135	0.5516
ESA-Skip-Gram	200	0.6287	0.6813	0.7322	0.4527	0.4278	0.5568
CBOW	200	–	0.3896	0.4674	0.3184	0.3191	0.3570
ESA-CBOW	200	–	0.4436	0.4850	0.3456	0.3210	0.3715
PV-DBOW	200	0.7656	0.6242	0.7039	0.5454	0.5237	0.6808
ESA-PV-DBOW	200	**0.8176**	0.7188	**0.7505**	0.5691	0.5423	**0.7120**
PV-DM	200	–	0.3592	0.4518	0.4982	0.4759	0.5538
ESA-PV-DM	200	–	0.3938	0.4785	0.5040	0.4986	0.5800
BOC	2874	0.6594	0.6228	0.5910	0.5559	0.5153	0.5342

TABLE III: F_1 scores of each category using different features and LibSVM.

Feature Category	ESA-PV-DBOW	PV-DBOW	BOW TFIDF	BOC	ESA
Internal Medicine	0.8436	0.8113	0.7930	0.7264	0.7355
Surgery	0.7725	0.7335	0.7286	0.6247	0.6066
Gynaecology	0.8926	0.8720	0.8234	0.6938	0.7351
Ear-Nose-Throat & Stomatology	0.7930	0.7309	0.7269	0.6745	0.5786
Paediatrics	0.7485	0.7012	0.7040	0.5778	0.6156

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 61572434), China Knowledge Centre for Engineering Sciences and Technology (No. CKCEST-2016-1-14) and Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP) (20130101110136).

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[2] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[3] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis." in *IJCAI*, vol. 7, 2007, pp. 1606–1611.

[4] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *ICML*, 2014, pp. 1188–1196.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013, pp. 3111–3119.

[6] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[7] G. Nestler, "Traditional chinese medicine," *Medical Clinics of North America*, vol. 86, no. 1, pp. 63–73, 2002.

[8] L. Wu, *Chinese traditional medicine and materia medical subject headings*. Beijing: Chinese Medical Ancient Books Publishing, 1996.