2019 18th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

On Herb Compatibility Rule of Insomnia Based on Machine Learning Approaches

1st Fang Hu
College of Information Engineering
Hubei University of Chinese Medicine
Wuhan, P.R. China
Department of Mathematics and Statistics
University of West Florida
Pensacola, U.S.A.
naomifang@hbtcm.edu.cn

2nd Liuhuan Li
College of Information Engineering
Hubei University of Chinese Medicine
Wuhan, P.R. China
1hLee@stmail.hbtcm.edu.cn

3rd Xiaoyu Huang
College of Basic Medicine
Hubei University of Chinese Medicine
Wuhan, P.R. China
huangxy1021@stmail.hbtcm.edu.cn

4th Panpan Huang *
College of Basic Medicine
Hubei University of Chinese Medicine
Wuhan, P.R. China
panpanhuang@aliyun.com

5th Liang Cheng
College of Information Engineering
Hubei University of Chinese Medicine
Wuhan, P.R. China
descarteshub@gmail.com

***Abstract*—Recent research in machine learning has led to significant progress in various research fields. Especially, the knowledge discovery using this method in Traditional Chinese Medicine (TCM) has been becoming a hot topic. In this paper, we studied on the herb compatibility rule of insomnia using some machine learning approaches. We have extracted insomnia data set with 807 samples from the real-world Electronic Medical Records (EMRs). After cleaning and selecting the theme data referring to the prescriptions and their herbs, we constructed the herb network analysis model using the theory of complex network. In order to explore the hidden relationships among the herbs, we trained each herb node in network to obtain the herb embeddings using the Skip-Gram model in word embedding theory. After acquiring the vocabulary of herbs with the formation of vectors, we calculated the similarity among any two herb embeddings, and clustered these herb embeddings into seven communities using the Spectral Clustering (SC) algorithm. The experimental results shed light on that the methodologies used in this paper can objectively and effectively discover the relationships among herbs, and reveal the herb compatibility and herb clusters for clinical treatment research of insomnia.**

***Index Terms*—Insomnia, Core Herb, Herb Community, Word Embedding, Spectral Clustering Algorithm**

I. Introduction

Machine learning, a subset of artificial intelligence and a kind of data-oriented approach, has attracted enormous attention from various domains [1]. Researchers already have proposed a huge number of algorithms and models referring to machine learning to discover the hidden relationships among entities from different research fields [2], [3]. The Traditional Chinese Medicine (TCM) data has the all characteristics of big data, especially, the relationships among the diseases, syndromes, symptoms, prescriptions, herbs, diagnosis, treatment, etc., are so complicated [4]. Insomnia, a chronic disease, affects daytime functioning, quality of life, and other mental and physical health outcomes. It is reported that there are about 16% of the population suffering from insomnia [5]. The clinical research denotes that TCM has been successfully applied in the treatment of insomnia. However, the evaluation criterion of TCM treatment of insomnia remains unexplored.

In recent decades, a great number of researchers exerted various machine learning approaches to discover the potential regulations for treating diseases using TCM. Du et al. [6] combined the partial least square and model tree to explore the hidden relationships among TCM data and get the precision predicted results. Hu et al. [7] discovered the core symptoms and symptom distribution rule of insomnia using the network analysis method. Li et al. [8] explored suitable pre-processing methods for the TCM clinical data based on prospective study on insomnia treated by syndrome differentiation. Zhong et al. [9] found the frequency of each herb and association rules among the herbs for insomnia using data mining methods. Because the features of TCM data is giant complex, how to use the effective methods to explore the TCM rules and support the clinical diagnosis and treatment is still a open question?

In this paper, we make attempts to explore the potential regularity of treating insomnia using TCM approach. By using the word embedding [10], we trained each herb node embedding in herb network, and constructed the specific herb vocabulary with the digital formation of vectors. Further, we clustered the herb communities through similarity calculation among any two herb embeddings. Finally, we obtained the results of core herbs and herb clusters, and summarized the herb compatibility rule of insomnia.

* Corresponding Author: panpanhuang@aliyun.com

2473-3636/19/$31.00 ©2019 IEEE
DOI 10.1109/DCABES48411.2019.00071

II. Insomnia Herb Network Construction

A. Data Set of Insomnia

The analysis data set of insomnia from the hospital information system in affiliated hospital Guo Yi Tang in Hubei University of Chinese Medicine. The regulation for selecting insomnia records is presented as follows:

(1) Inclusion Criteria

The patients are diagnosed with typical symptoms of insomnia; the sleep disorder is the main symptom of patients and the other symptoms are secondary to insomnia; the age range of patients is $14 \sim 70$; the course of insomnia is between 1 month and 30 years.

(2) Exclusion Criteria

Non-partners, including those who are unable to adhere to treatment or affect data collection and efficacy evaluation; pregnant women or terminally ill patients.

Based on the aforementioned criteria, we have extracted 807 effective outpatient Electronic Medical Records (EMRs). Through analyzing the theme data used in this manuscript, we cleaned and selected some significant features including syndromes, prescriptions and their corresponding herbs, and then formed the analysis data set of insomnia.

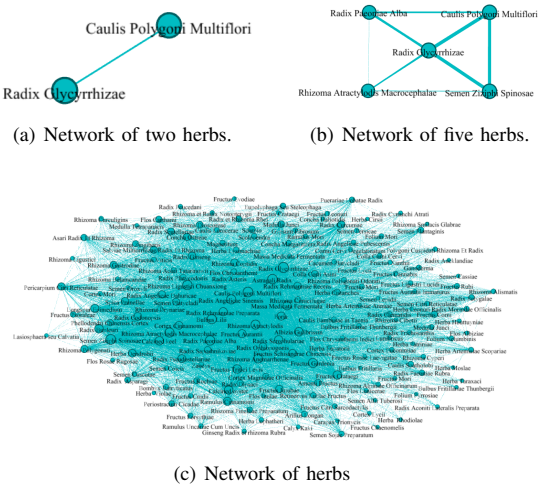
B. Herb Network Model

Based on the theory of complex network [11], [12], we constructed the insomnia herb network $G(V, E)$, where V is the node set of herbs and E denotes the edge set among any two herbs. The regular of herb network construction is as follows: we take each herb in the records as a node in the network; consider the connection among any two herb co-occurred in the same prescription as an edge in the network; denote the weight of an edge as the co-occurrence times of any two herbs.

According to the aforementioned rule, we present the construction process of insomnia herb network in Figure 1. As in Figure 1(a), we constructed network with two herb nodes $Caulis\ Polygoni\ Multiflori$ and $Radix\ Glycyrrhizae$, and an edge denoting these two herbs co-occurred in the same prescription. In development, the other three herb nodes $Radix\ Paeoniae\ Alba$, $Rhizoma\ Atractylodis\ Macrocephalae$, and $Semen\ Ziziphi\ Spinosae$, and their corresponding weighted edges were added into this network in Figure 1(b). Finally, we acquired a weighted and undirected herb network of insomnia with 238 nodes and $16,376$ edges in Figure 1(c). Based on this Figure, the size of one node represents the importance of this herb. If a herb has more relationships (edges) with other herbs, in other word, it denotes that this herb is significant for treating insomnia.

III. Methodologies

The summary of the methodologies for insomnia data processing is outlined in Figure 2. We divided the data processing into three sections: data preparation, data training, and data clustering.



(a) Network of two herbs. (b) Network of five herbs.

(c) Network of herbs

Fig. 1. The construction process of herb network.

In the first section, we got the original EMRs data set S from the hospital information system; cleaned and selected the herb information from S; then, we constructed the herb information table HT; after encoding each herb, the list of ordered pair herb-herb HH was acquired.

In the second section, we constructed the herb network based on HH; calculated the transition probability for each herb node and normalized the probability, using which we acquired the walk sequence P of herb nodes; after training P using Skip-Gram model, the herb embeddings E were obtained.

In the third section, we calculated the weight matrix W, i.e., the similarity matrix, based on the herb embeddings E; Through getting the degree matrix D and the Laplacian matrix L, we obtained the eigenvector matrix F; after clustering F using K-means algorithm, the herb communities C were acquired.

IV. Training the Herb Embeddings

Based on the matrix of herb network, we used the Skip-Gram model [13] to train the insomnia herb embeddings (also called herb vectors). We first built a vocabulary of 238 insomnia herb terms. We represented an input herb term like $Radix\ Glycyrrhizae$ as a one-hot vector. This vector have 238 components (one for every herb in our vocabulary) and we placed a "1" in the position corresponding to the herb $Radix\ Glycyrrhizae$, and 0s in all of the other positions. The output of the network was a single vector (also with 238 components) containing, for each herb in our vocabulary, the probability that a randomly selected nearby herb was that vocabulary herb. The neural network model for training the herb embeddings was in Figure 3. In this model, we set the input layer as the 238 one-hot herb vectors; the number of the neurons in hidden layer as 128; the activation function in output layer as softmax function. When we evaluated the trained

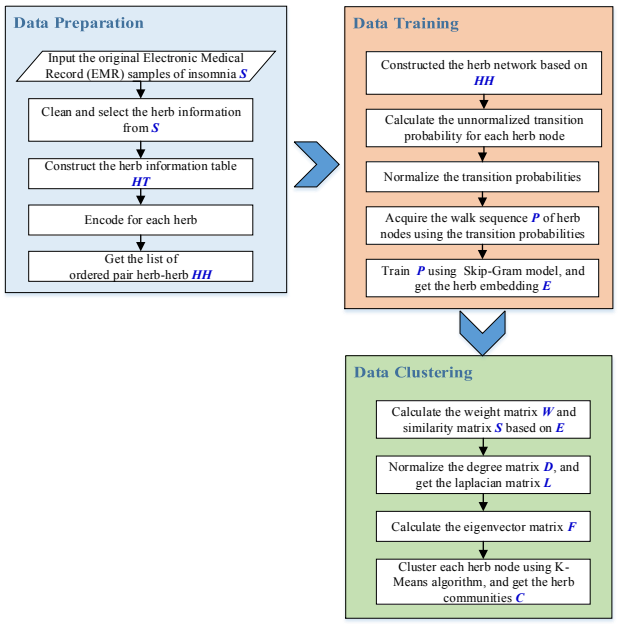


Fig. 2. Flowchart of data processing.

network on an input herb one-hot vector, the output vector actually was a probability distribution (i.e., a bunch of floating point values, not a one-hot vector). Therefore, we got the probabilities of the herbs, such as *Semen Ziziphi Spinosae*, *Rhizoma Atratylodis Macrocephalae*, *Astragali Radix*, *Radix Paeoniae Alba*, etc., appeared nearby the herb *Radix Glycyrrhizae*. After training the model in Figure 3,

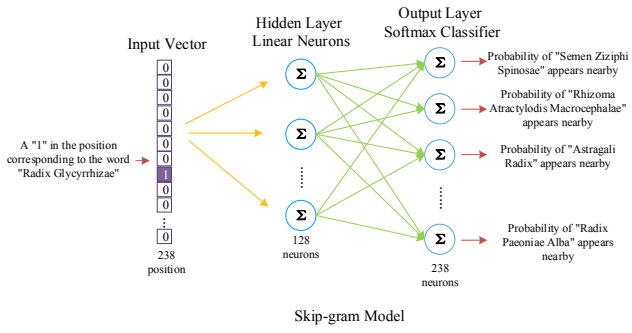


Fig. 3. Skip-Gram Model of herbs.

we acquired the weight matrix (i.e., the herb embeddings with 128 features) in hidden layer. This weight matrix has 238 rows (one for each herb in our vocabulary) and 128 columns (one for every hidden neuron). The herb embedding lookup table is obtained from the weight matrix in hidden layer and we show it in Figure 4.

V. Clustering the Herb Embeddings

In order to find the rule of herb compatibility and herb clusters of insomnia, we used the Spectral Clustering (SC)

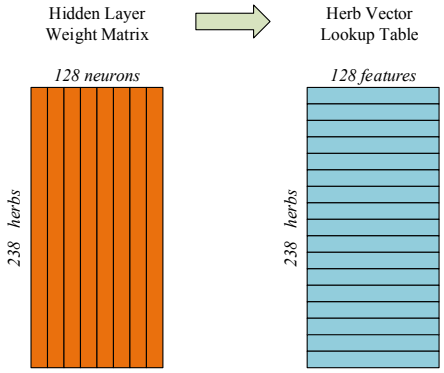


Fig. 4. Representation of Herb Embeddings.

algorithm [14], [15], one of the representative community detection algorithms in complex networks, to divide the herb network with 238 nodes and 16,376 edges into the real communities. A community is constituted of one group or cluster of nodes within which the links between nodes are densely connected to each other but between which they are sparsely connected with other communities [16].

We calculated the similarity values among any two herb embeddings and divided the herbs with high similarity values into the same community. The clustering process is as follows: we constructed the weight matrix W (i.e., similarity matrix) through calculating the special distance between arbitrary two herb nodes v_i and v_j; got the degree matrix D; calculated the Laplacian matrix $L = D - W$ and obtained the normalized Laplacian matrix L'; found the first k minimum eigenvalues and their corresponding eigenvectors of L', and constructed the eigenmatrix F using these eigenvectors; clustered the eigenmatrix F using K-means algorithm and acquired the herb clusters of insomnia. We show the herb community structure in Figure 5. As shown in Figure 5, this herb network (Figure 1)

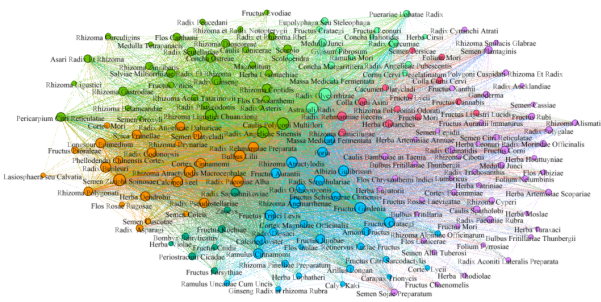


Fig. 5. Herb communities.

is split into seven communities using the SC algorithm. Based on the regularity of network construction, the various sizes of nodes denote the different importance of herbs of insomnia, a fact that the size of one node is bigger than other nodes representing this herb is more important. The sizes of edges represent the co-occurrence frequencies of any two herbs in

prescriptions. The clustering result revealed that classic herb clusters of insomnia and was shown as follows:

Community 1 (colored by green): the herbs *Radix Angelicae Sinensis*, *Rhizoma Ligustici*, *Radix Platycodonis*, *Flos Chrysanthemi*, *Fructus Viticis*, etc., have been divided into a community with the core herb *Caulis Polygoni Multiflori*.

Community 2 (colored by light green): the herbs *Astragali Radix*, *Radix Curcumae*, *Medulla Junci*, *Fructus Leonuri*, etc., have been divided into a community with the core herb *Radix Glycyrrhizae*.

Community 3 (colored by pink): the herbs *Colla Corii Asini*, *Rhizoma Cimicifugae*, *Rhizoma Polygonati Odorati*, *Semen Persicae*, *Cacumen Platycladi* etc., have been divided into a community with the core herb *Radix Rehmanniae Recens*.

Community 4 (colored by purple): the herbs *Radix Polygalae*, *Radix Clematidis*, *Semen Citri Reticulatae*, *Cortex Eucommiae*, *Herba Taraxaci*, etc., have been divided into a community with the core herb *Fructus Corni*.

Community 5 (colored by blue): the herbs *Albizia Gulibrissin*, *Fructus Aurantii*, *Fructus Tritici Levis*, *Arillus Longan*, *Radix Scrophulariae*, etc., have been divided into a community with the core herb *Fructus Gardenia*.

Community 6 (colored by dark green): the herbs *Fructus Cnidii*, *Fructus Crataegi*, *Fructus Forsythiae*, *Zaocys*, *dizziness*, etc., have been divided into a community with the core herb *Radix Saposhnikoviae*.

Community 7 (colored by orange): the herbs *Cortex Cinnamomi*, *Semen Cuscutae*, *Radix Asparagi*, *Rhizoma Polygonati*, *Herba Dendrobii*, etc., have been divided into a community with the core herb *Semen Ziziphi Spinosae*.

VI. Conclusions

Recently, the knowledge discovery in TCM using machine learning approaches has been a hot research topic. In this study, we took the insomnia as the research instance, and explored the herb compatibility rule using some machine learning approaches, especially, the word embedding. We constructed the herb network to reflect the hidden relationships among the herbs, a fact that can intuitively and clearly depict the compatibility relationships among herbs commonly used in the treatment of insomnia. Then, we trained the herb vocabulary using the word embedding. After the clustering analysis of herb embeddings using SC algorithm, we identified the significant herbs, which can standardize, structure, and scientize the invisible knowledge of common herbs. Finally, we acquired the insomnia herb communities, which can reveal the herb compatibility relationship. We show that the methodologies can effectively and accurately discover the hidden relationship among herbs for treating insomnia. The herb compatibility rule summarized from the experimental results can provide new prescription ideas and treatment schemes for clinical treatment of insomnia, and also provide guidance and reference for clinical work. The trained insomnia herb embeddings can be supplied for the further research as a basis data base. In development, we can use the similar approaches to explore the herb compatibility rule for other disease treatment.

Acknowledgement

We acknowledge the funding support by the National Natural Science Foundation of China (81874414), and the Natural Science Foundation of Hubei Province (2018CFB259).

References

[1] M. Allamanis, E. T. Barr, P. Devanbu, and C. Sutton, "A survey of machine learning for big code and naturalness," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–37, 2018.

[2] Q. Su, Y. Zhu, Y. Jia, P. Li, F. Hu, and X. Xu, "Sedimentary environment analysis by grain-size data based on mini batch k-means algorithm," *Geofluids*, vol. 2018, p. 8519695, 2018.

[3] F. Hu, M. Wang, Y. Zhu, J. Liu, and Y. Jia, "A time simulated annealing-back propagation algorithm and its application in disease prediction," *Modern Physics Letters B*, vol. 32, no. 25, p. 1850303, 2018.

[4] X. Zhou, Y. Peng, and B. Liu, "Text mining for traditional chinese medical knowledge discovery: a survey," *Journal of biomedical informatics*, vol. 43, no. 4, pp. 650–660, 2010.

[5] S. E. Emert, J. Tutek, and K. L. Lichstein, "Associations between sleep disturbances, personality, and trait emotional intelligence," *Personality and Individual Differences*, vol. 107, pp. 195–200, 2017.

[6] D. U. Jian-Qiang, Y. U. Fang, B. Nie, Z. P. Zhu, L. Liu, and S. O. Computy, "Analysis of tcm data based on partial least square and model tree," *Chinese Journal of New Drugs*, vol. 26, no. 17, pp. 1997–2002, 2017.

[7] H. Fang, Y. Qiao, G. Xie, Y. Zhu, Y. Jia, and P. P. Huang, "Symptom distribution regulation of core symptoms in insomnia based on informapsa algorithm," in *2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES)*, 2017.

[8] L. X. LI, Y. Liu, N. Wang, J. A. Hou, H. S. Wang, Z. X. Zhou, S. R. Zhang, Y. B. Liu, and Y. L. HE, "Study on pre-processing methods of clinical data from tcm individual treatment of insomnia based on syndrome differentiation," *Chinese Journal of Information on Traditional Chinese Medicine*, vol. 24, no. 12, pp. 92–96, 2017.

[9] S. Weng and N. Zhou, "Analysis on zhong yi-tang's medication rule in prescriptions for insomnia based on data mining method," *Journal of Zhejiang Chinese Medical University*, no. 8, pp. 595–597, 2015.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[11] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[12] F. Hu, Y. Zhu, J. Liu, and Y. Jia, "Computing communities in complex networks using the dirichlet processing gaussian mixture model with spectral clustering," *Physics Letters A*, vol. 383, no. 9, pp. 813–824, 2019.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[14] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak mathematical journal*, vol. 23, no. 2, pp. 298–305, 1973.

[15] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[16] M. E. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical review E*, vol. 74, no. 3, p. 036104, 2006.