Knowledge discovery in Chinese herbal medicine: a machine learning perspective

Nan Liang[1], *Qing* Liang[2,*], and *Fenglei* Ji[3]

[1]Center for Computer Fundamental Education, Jilin University, Changchun 130012, P.R. China
[2]College of Life Sciences, Jilin University, Changchun 130012, P.R. China
[3]College of Communication Engineering, Jilin University, Changchun 130012, P.R. China

Abstract. Traditional Chinese Medicine (TCM) has attracted more and more attention due to its remarkable effects on treating diseases, and Chinese herbal medicine (CHM) is an important partition of TCM, rich in natural active ingredients. Researchers are trying multiple analytical methods to dig out more valuable information about CHM and reveal the principle of TCM. Machine learning is playing an important role in the studies. Knowledge discovery of CHM using machine learning mainly includes quality control of CHM, network pharmacology in CHM, and medical prescriptions composed by CHM, aiming to understand TCM better, provide more efficiency methods in the production of CHM and find novel treatment of disease not curable nowadays. In this paper, we summarized the basic idea of frequently used classification and clustering machine learning algorithms, introduced pre-processing algorithms commonly used to simplify and accelerate machine learning procedure, presented current status of machine learning algorithms' applications in knowledge discovery of CHM, discussed challenges and future trends of machine learning's application in CHM. It is believed that the paper provides a valuable insight for the starters trying to apply machine learning in the study of CHM and catch up the recent status of related researches.

1 Introduction

Traditional Chinese Medicine (TCM) is a valuable medical method in China based on experience accumulated in thousands of years. TCM is a large and complex system, the ingredients of Chinese herbal medicine (CHM) are very complicated, and the mechanism of TCM treatment is not clear yet [1].

In the era of big data, how to dig out valuable information from the tremendous data is important, so is in the domain of TCM. Machine learning is introduced as an effective information extraction tool from raw and heterogeneous data, and proved to be a successful method in many fields. Many researches have been trying to figure out how to combine machine learning and TCM together, to reveal the principle of TCM to the world. In this paper, we reviewed the machine learning methods commonly used in CHM researches,

[*] Corresponding author: liangqing@jlu.edu.cn

introduced main applications of machine learning in CHM, and discussed challenges and future trends of machine learning applied in CHM.

2 Machine learning algorithms in Chinese herbal medicine

Support vector machine (SVM) has been applied in various domains because of good generalization performance. SVM's goal is to find the 'optimal hyperplane' to classify training data, realizing the best learning ability with limited input samples.

Neural Network (NN) uses layers of neurons to process distributed parallel information, build a stable and precise prediction network. The feature of NN includes good non-linear mapping capabilities and able to recognize noisy input data.

As representative of lazy learning, K-Nearest Neighbor (KNN) is the easiest machine learning algorithm to interpret and understand. KNN's training process is simpler, efficiency is higher, and performance is compatible with other algorithms [2]. KNN is computationally expensive due to calculation of pairwise distance.

Naïve Bayes (NB) is fast, capable of processing large dataset, tolerant to random noise [3], and needs less training data. NB assumes each attribute is independent to each other. NB obtains the classification results based on the calculation of conditional probabilities.

k-means is a simple and effective clustering algorithm, after choosing k centroids in the dataset, every data is assigned to its closest centroid's cluster. The convergence of k-means is to a local optimum.

Hierarchical clustering is a group of clusters organized as a tree showing the merging process. Bottom-up and top-down are two kinds of hierarchical clustering process. How clusters are divided is based on the dissimilarity between the set of objects [4].

Besides, principal component analysis (PCA) and normalization are popular pre-processing algorithms in machine learning. PCA simplify complex datasets by reducing the dimensionality Normalization speeds up the learning phase by mapping data into certain range.

3 Knowledge discovery in Chinese herbal medicine

Typical application of machine learning in knowledge discovery of CHM includes: quality control of CHM, network pharmacology researches in CHM and knowledge discovery in CHM prescriptions.

3.1 Quality control of Chinese herbal medicine

Some CHM have similar appearance, discriminating them manually is difficult. Misidentification could cause shortage of medicine. Therefore, methods to guarantee quality of CHM are important. In the discrimination and quality evaluation of CHM, spectroscopy and chromatography are two mainly used methods combining with machine learning algorithms.

Infrared spectroscopy and near-infrared spectroscopy (NIR) are the most commonly used spectroscopy technology to get the data inputting to machine learning models. In [5], researchers combined SVM with Fourier transform infrared spectroscopy to identify of main root, rhizome and fibrous root of Panax notoginseng. NIR could discriminate geographical origin of herbal medicine when utilizing machine learning methods. In [6], after NIR, BPNN, KNN and SVM were constructed as classification model to predict the origin of CHM. Results indicated SVM model gave the best discrimination rate.

For chromatography, high-pressure liquid chromatography (HPLC) and ultra-performance liquid chromatography (UPLC) were commonly used in quality control of CHM. Hierarchical clustering analysis (HCA) was the most frequently used clustering algorithm combined with chromatography, and the application were concentrated on discrimination of CHM. In [7], authors conducted HPLC to simultaneously quantify artemisinin and six synergistic components in CHM. Then based on HPLC data, HCA could identify different parts and plucking times of medicine. Meng et al. applied HCA based on UPLC fingerprint to efficiently identify three TCM drugs in the Paeoniaceae family [8].

3.2 Network pharmacology researches in Chinese herbal medicine

Machine learning methods are widely used to find potential inhibitors and lead drugs in CHM to proceed the process of new drug development. Researchers also are interested in revealing synergistic mechanism of TCM. SVM and NB were the most commonly used machine learning methods in the related researches.

SVM were used to predict PepT1 (an essential target for drug) substrates from TCM database, the data in the train set was from previous researches and Drugbank. Then pharmacophore and docking model screened potential activate gene expression of PepT1s agonists from TCM database to reveal the synergistic mechanism of TCMs [9]. To improve mechanistic understanding of TCM, [10] utilized NB and HCA to explore global mapping of CHM compounds based on the therapeutic action classes.

3.3 Knowledge discovery for prescriptions of Chinese herbal medicine

In knowledge discovery for prescriptions of CHM, researches concentrated in two directions: finding CHM prescriptions to treat certain diseases, and revealing the mechanism of TCM based on the properties of CHM.

NN and SVM were often used in finding CHM prescriptions to treat certain diseases. In [11], researchers applied NN and k-means to identify TCM patterns based on dataset of breast cancer patients, results could help to identify effective CHM treatments in clinical practice. SVM was chosen to combine with network pharmacology to find novel TCM formula toward Huntington’s disease [12].

Plenty of researches focused on explanation of systematical mechanisms of TCM prescriptions. NB was utilized in [13] to predict component–target spectrum for all the compounds in Wenxin Keli. The overlap of component–target spectrum and disease-target spectrum of cardiac arrhythmia explained the efficacy of Wenxin Keli.

4 Discussion

Though machine learning has been widely applied in the researches of CHM, and performances turned out to be well, challenges and issues still emerges as the researches going on.

4.1 Challenges and Issues

CHM data could influence the performance of algorithms significantly. Larger datasets often lead to better prediction of machine learning algorithm, but they also need better storage technology and challenge to computing infrastructure. Data sources of CHM usually come from different databases with heterogeneous structures, thus fusion of data is complex. Raw datasets always contain missing and duplicate data, a better data pre-processing method is needed before prediction.

Choosing appropriate machine learning algorithm is also widely discussed, lots of researches compared different algorithms' performance and chose the better one in certain problems [2]. Pros and cons of commonly used machine learning algorithms are shown in Table1. The hyperparameters, such as value k of KNN, kernel function of SVM, need to be optimized before modeling. Because data in training set is often less than enough, the evaluation of algorithms' performance tended to be of high variance. Generalization ability of model also needs to be paid attention because sometimes the model claimed better prediction accuracy only fitted for the certain research.

Table 1. Comparison of advantages and disadvantages of different machine learning algorithms in classification.

Machine learning algorithm	**Advantages**	**Disadvantages**
SVM	More accurate More robust Better ability of dealing with nonlinear data	Different kernel for different application Mainly for binary classification
NN	Ability to deal with complex and nonlinear problems High accuracy Failure tolerance	Black box phenomenon
KNN	Simple and fast	Sensitive to noise
NB	Faster and more accurate when dataset is large Easy to understand Tolerant to noise	Result may not be right when variables are dependent to each other
k-means clustering	Simpler principle Faster convergence speed	Local optimum
Hierarchical clustering	Visualize a structure in the form of a taxonomy More stable	High computational complexity

4.2 Future Trends

Machine learning algorithms are of various features, researchers have tried to combine multiple algorithms together to take advantage of their pros, and some studies showed the performances turning out to be well. In identification and classification of CHM, Xiyang et al. presented the results that ensemble learning integrated with SVM and PNN could achieve a better accuracy than single classifier [14]. As a representative of ensemble learning, random forest also has been introduced in constructing model for quality control in CHM.

As machine learning algorithms require large amount of labeled data, it needs lots of manual work and specialized knowledge. Deep learning could point out how a model adjusting its internal parameters with less human labor. Thus deep learning algorithms such as deep belief network, convolutional neutral network are applied in the knowledge discovery of CHM [2].

Besides deep learning, other machine learning methods, such as spectral clustering and extreme learning machine, were also applied in knowledge discovery in CHM, and results turned well. For example, in classification of TCM of different ages, extreme learning machine reached best performance compared with SVM, BPNN and random forest [15].

Because of the heterogeneity of data source, novel data fusion methods are also needed to be addressed. Besides, the models were usually evaluated by the indicators such as accuracy, sensitivity. More comprehensive methods need to be addressed for evaluating the performance of model.

Because CHM and their formulas have complex compositions, recent researches tried to find novel prescriptions in treating diseases that are hard to treat or not curable, such as cancer, insomnia and so on [11].

5 Conclusion

As a unique medical method, more and more researches have taken TCM into consideration to reveal complex principle and take better advantages of it. CHM is an important part of TCM. Machine learning is one important method to deepen the understanding of CHM. As a result, it's believed that machine learning could make great contribution to the development of TCM.

This research was supported by the Fundamental Research Funds for the Central Universities.

References

1. P. Gu and H. Chen, *Briefings in Bioinformatics*, **15**, 984-1003, (2013).
2. Z. Chen, Y. Cao, S. He and Y. Qiao, *Chin. Med.*, **13**, 12, (2018).
3. S. Tian, J. Wang, Y. Li, X. Xu and T. Hou, *Mol. Pharm.*, **9**, 2875-2886, (2012).
4. M. K. Rafsanjani, Z. A. Varzaneh and N. E. Chukanlo, *Journal of Mathematics and Computer Science*, **5**, 229 - 240, (2012)
5. Y. Li, J. Zhang, H. Jin, Y. Wang and J. Zhang, *Guang pu xue yu guang pu fen xi = Guang pu*, **39**, 103-108, (2019).
6. Y. Yang, Y. Wu, W. Li, X. Liu, J. Zheng, W. Zhang and Y. Chen, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, **191**, 233-240, (2018).
7. F. Qiu, S. Wu, X. Lu, C. Zhang, J. Li, M. Gong and M. Wang, *Ind Crop Prod*, **118**, 131-141, (2018).
8. M. Lu, Q. Hu, Y. Zhang, Y. Zhai, Y. Zhou and J. Jiang, *Biochem. Syst. Ecol.*, **83**, 121-129, (2019).
9. L. Qiao, Y.-K. Chen, G.-G. Luo, F. Lu, S. J. Liu, G.-Y. Li and Y. Zhang, *Zhongguo Zhong yao za zhi = Zhongguo zhongyao zazhi = China journal of Chinese materia medica*, **42**, 2146-2151, (2017).
10. S. Z. Mohamad Zobir, F. Mohd Fauzi, S. Liggi, G. Drakakis, X. Fu, T.-P. Fan and A. Bender, *Evid. Based Complement. Alternat. Med.*, **2016**, 2106465, (2016).
11. W.T. Huang, H.H. Hung, Y.W. Kao, S.C. Ou, Y.C. Lin, W.Z. Cheng, Z.R. Yen, J. Li, M. Chen, B.C. Shia and S.T. Huang, *Front. Pharmacol.*, **11**, 670, (2020).
12. W. Dai, H.-Y. Chen and C. Y.-C. Chen, *Evid. Based Complement. Alternat. Med.*, **2018**, 6020197, (2018).
13. T. Wang, M. Lu, Q. Du, X. Yao, P. Zhang, X. Chen, W. Xie, Z. Li, Y. Ma and Y. Zhu, *Mol Biosyst*, **13**, 1018-1030, (2017).
14. X. Sun, L. Liu, Z. Wang, J. Miao, Y. Wang, Z. Luo and G. Li, *Sensors and Actuators A: Physical*, **266**, 135-144, (2017).
15. Q. Shi, *International Journal of Electrochemical Science*, **13**, 11359-11374, (2018).