Discovery of Regularities in the Use of Herbs in Traditional Chinese Medicine Prescriptions

Nevin L. Zhang[1], Runsun Zhang[2], and Tao Chen[3]

[1] Department of Computer Science & Engineering,
The Hong Kong University of Science & Technology,
Clear Water Bay, Kowloon, Hong Kong
lzhang@cse.ust.hk
[2] Guanganmen Hospital,
Chinese Academy of Chinese Medical Sciences,
Beijing, China
[3] EMC Labs China
Beijing, China
tao.chen2@emc.com

Abstract. Traditional Chinese medicine (TCM) is a discipline with its own distinct methodologies and philosophical principles. The main method of treatment in TCM is to use herb prescriptions. Typically, a number of herbs are combined to form a formula and different formulae are prescribed for different patients. Regularities on the mixture of herbs in the prescriptions are important for both clinical treatment and novel patent medicine development. In this study, we analyze TCM formula data using latent tree (LT) models. Interesting regularities are discovered. Those regularities are of interest to students of TCM as well as pharmaceutical companies that manufacture medicine using Chinese herbs.

Keywords: Herb regularities, latent tree model, traditional Chinese medicine prescription.

1 Introduction

Traditional Chinese medicine (TCM) is a discipline with its own distinct methodologies and philosophical principles [1]. It has successfully prevented the Chinese and East Asia people from serious diseases for thousands of years. As one of the oldest healing systems, TCM includes the therapies like herbal medicine, acupuncture, moxibustion, massage, food therapy, and physical exercise [2]. They can be practically used for various diseases treatment [3]. The herbal medicine, generally called formula, is one of the most important TCM therapies. It tries to acquire maximal therapeutic efficacy with minimal adverse effects. Typically, a formula consists of several medicinal herbs or minerals (we will use the word "herb" to refer to medicinal materials in formula). Different components in a formula have different 'roles' for disease treatment [4]. Clinical herb prescription is

L. Cao et al. (Eds.): PAKDD 2011 Workshops, LNAI 7104, pp. 353–360, 2012.
© Springer-Verlag Berlin Heidelberg 2012

a complicated and flexible procedure that integrates the knowledge of syndrome differentiation (i.e., TCM diagnosis), TCM herb and formula theories, treatment principles, and empirical herb prescription knowledge inherited from the ancient literatures and acquired through individual experiences. In contrast to the modern drug therapies that often adhere to the common and operational clinical guidelines, TCM physicians emphasize more on individuality when prescribing formulae in TCM clinical practices. The formulae prescribed for different patients are almost never the same. A large amount of formula data, along with other clinical information, has been accumulated over the years. To manage all the data, Zhou *et al.* [5] have developed a clinical data warehouse.

In this study, we are interested in discovering the regularities on herb combination from large-scale clinical herb prescription databases. Several data mining methods have been used for the purpose before [6]. The results are not satisfying. Take association rules as an example. It is the most commonly used method. A key drawback is that it produces a large number of rules, often in the thousands. Clinical researchers have to painstakingly go through all the rules to get the final discoveries. This takes a lot of time and efforts. Moreover, association rules are concerned with only co-occurrence patterns, while it is far more interesting to analyze the complicate interactions, e.g. synergy, mutual detoxification and mutual inhibition, among the herbs in the clinical formulae. Furthermore, the co-occurrence frequency-based methods can not detect the negative dependence between herbs, which is important for clinical practices.

Zhang *et al.* [7,8] have studied the discovery of TCM diagnosis knowledge from the clinical data using a new class of statistical methods called latent tree (LT) models. The models enable one to discover the latent structures based on local dependences between the manifestation variables [9]. Technically, an LT model is a tree-structured Bayesian network where variables at leaf nodes are observed and are hence called "manifest variables", whereas variables at internal nodes are hidden and hence are called "latent variables". All variables are assumed to be discrete. Arrows represent direct probabilistic dependence.

In this study, we use LT models to analyze TCM formula data and thereby reveal the underlying latent structures. In particular, we analyzed the herb prescription data for patients in a condition known in TCM as "disharmony between liver and spleen (DBLS) syndrome". The data were extracted from a data warehouse [5], and the prescriptions were made by senior and well known TCM experts. The analysis has revealed some clinically useful regularities. Common herb combinations and their modifications for the treatment of DBLS were discovered. The results are useful for the students of TCM as well as pharmaceutical companies that manufacture medicine using Chinese herbs.

The rest of this paper is organized as follows. We introduce the latent tree models and the clinical data in turn in section 2 and 3. The results are presented in section 4. Finally, we discuss the clinical significance of the study and the future work in section 5.

2 Latent Tree Models

A *latent tree (LT) model* is a Bayesian network where (1) the network structure is a rooted tree; (2) the internal nodes represent latent variables and the leaf nodes represent manifest variables; and (3) all the variables are categorical. *Latent class (LC) models* are LT models with a single latent node. The terms variable and node are interchangeable throughout this paper.

Figure 1 (a) shows the structure of an LT model. In the model, there is an arrow from variable Y_1 to variable Y_2. This means that Y_2 depends on Y_1 directly. The dependence is quantified by a conditional distribution $P(Y_2|Y_1)$, which gives a distribution for Y_2 for each value of Y_1. All these distributions forms the *parameters* of a latent tree model. We write an LT model as a pair $M = (m, \theta)$, where θ is the collection of parameters. The first component m consists of the variables, the cardinalities of the variables, and the model structure. We sometimes refer to m also as an LT model.

Assume that there is a collection $\mathcal{D}$ of i.i.d. samples on manifest variables generated by an unknown LT model. By *LT model learning* we mean the effort to reconstruct the generative model from the data. The search-based approach aims at maximizing a scoring function. The BIC score [10] of a model m is:

$$BIC(m|\mathcal{D}) = \max_{\theta} \log P(\mathcal{D}|m, \theta) - \frac{d(m)}{2} \log N,$$

where $d(m)$ is *dimension*, i.e., the number of independent parameters of the model, and N is the sample size. The first term on the right hand side is known as the *maximized loglikelihood of* m. It measures how well model m fits the data $\mathcal{D}$. The second term is a penalty term for model complexity.

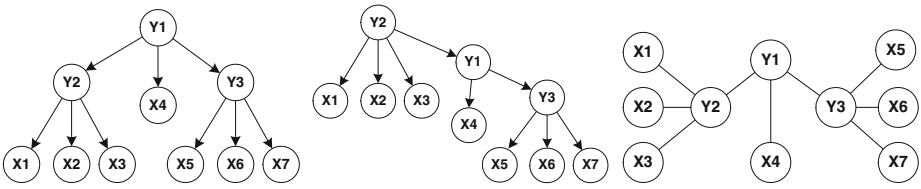
Consider two LT models m and m' that share the same manifest variables $X_1, X_2, \ldots, X_n$. We say that m *includes* m' if for any parameter value θ' of m', there exists parameter value θ of m such that

$$P(X_1, \ldots, X_n|m, \theta) = P(X_1, \ldots, X_n|m', \theta').$$

When this is the case, m can represent any distributions over the manifest variables that m' can. If m includes m' and vice versa, we say that m and m' are *marginally equivalent*. Marginally equivalent models are *equivalent* if they have the same number of independent parameters. It is impossible to distinguish between equivalent models based on data if penalized likelihood score is used for model selection.

Let Y_1 be the root of a latent tree model m. Suppose Y_2 is a child of Y_1 and it is also a latent node. Define another latent tree model m' by reversing the arrow $Y_1 \rightarrow Y_2$. Variable Y_2 becomes the root in the new model. The operation is called *root walking*; the root has walked from Y_1 to Y_2. The model m' in Figure 1 (b) is the model obtained by walking the root from Y_1 to Y_2 in model m.

It has been shown that root walking leads to equivalent models [11]. Therefore, the root and edge orientations of an LT model cannot be determined from data. We can only learn *unrooted LT models*, which are LT models with all directions



(a) m (rooted) (c) m' (after root walking) (c) m (unrooted)

Fig. 1. Rooted latent tree models, latent tree model obtained by root walking, and unrooted latent tree model. The Xs are manifest variables and the Ys are latent variables.

on the edges dropped. An example of an unrooted LT model is given in Figure 1 (c).

An unrooted LT model represents an equivalent class of LT models. Members of the class are obtained by rooting the model at various nodes. Semantically it is a Markov random field over an undirected tree. The leaf nodes are observed while the interior nodes are latent. Marginal equivalence and equivalence can be defined for unrooted LT models in the same way as for rooted models. Henceforth LT models always mean unrooted LT models in this paper unless it is explicitly stated otherwise.

3 The Clinical Data

TCM differentiates between different patients based mainly on their symptoms. The classification is known as *syndrome*, or *pattern*. Patients with different diseases might have the same syndrome manifestations, and the same disease might have different syndrome manifestation at different stages. TCM treatment is mainly targeted at the syndromes rather than the diseases.

DBLS is a general syndrome that can manifest in many chronic diseases, such as chronic gastritis, fatty liver, infertility and liver cirrhosis. The main symptoms of DBLS are "*irritability, mental-emotional depression, chest, rib-side and abdominal distention or pain, premenstrual breast distention and pain, painful menstruation, fatigue, reduced food intake, stomach and epigastric distention and fullness after eating*", etc [3].

The DBLS syndrome has complicated symptoms and hierarchical pathological structures. There are a large variety of herb prescriptions [12]. The data set we analyzed consists of 1,287 clinical formulae for the DBLS syndrome. Most of them are prescribed by famous senior TCM physicians from the best TCM hospitals in Beijing. The formula data contain totally 367 distinct herbs. Each prescription contains 15 herbs on average. The top 5 most frequently used herbs are *indian bread* (1,107), *stir-frying largehead atractylodes rhizome* (1,011), *Chinese thorowax root* (974), *white peony root* (880) and *Chinese angelica* (719). This means that most formulae for DBLS syndrome have the above five herbs as

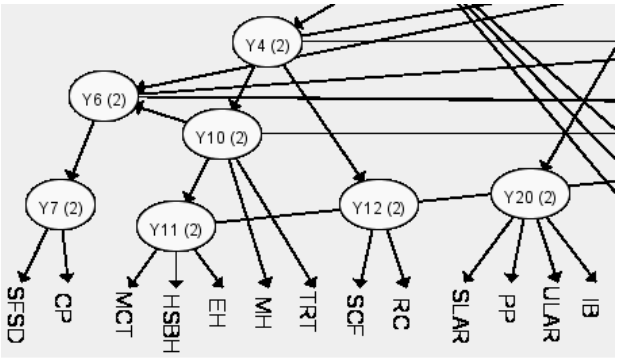


Fig. 2. Part of the LT model for the DBLS herb prescription data. The herb variables are at the leaf nodes and the latent variables are at the internal nodes. The numbers in parentheses are the numbers of states of the latent variables. The full names of the herbs are provided in Table 1.

ingredients. These 5 herbs are the core ingredients of a famous ancient formula known as *Xiao Yao San*. Most of the other herbs are in relatively low frequency.

Due to the intrinsic complexity in learning LT models, we reduced the number of variables and only the top 102 frequent herbs were included in the analysis. Each of those herbs appeared in at least 30 formulae. The herb variables were converted into binary variables, taking value '0' or '1'. The final data set is represented as a table with 102 attributes and 1,287 records.

4 The Results

We conducted LT analysis on the data set. The analysis was run on a 20-node cluster, where every node is of hard configuration: 2 x dual core AMD Opteron 2216 (2.4GHz) processors. The process took 156.9 hours to finish. The BIC score of the resultant model is -32,739. The resulting LT model has 38 latent variables. A portion of the model structure is shown in Fig. 2. This means that our analysis has identified 38 latent factors from the DBLS data set.

Each of the latent variables has a number of states. For example, the variable $Y1$ has 2 states. This means that we have grouped the data set into 2 clusters according to the latent factor $Y1$. The variable $Y11$ also has 2 states. This means that we have grouped the data set in another way into 2 clusters. Thus, we have simultaneously clustered the data set in multiple ways.

In the following, we examine several latent variables and their states to demonstrate the interestingness of the results.

The latent variable $Y7$ is connected to herb variables SFSD and CP, and latent variable $Y6$. $Y7$ has two states $S0$ and $S1$, which indicates that it represents a partition of the prescriptions into two classes. In the following, we discuss the meanings of this partition and hence the variable $Y7$ itself.

Table 1. The full name of some of the herbs

SFSD: stir-fry flying squirrel's droppings	SCF: szechwan chinaberry fruit
MCT: medicinal cyathula toot	RC: rhizoma corydalis
HSBH: hirsute shiny bugleweed herb	SLAR: stir-fry largehead atractylodes rhizome
EH: epimedium herb	PP: purified pinellia [tuber]
MH: motherwort herb	ULAR: uncooked largehead atractylodes rhizome
TRT: turmeric root tuber	IB: indian bread

Table 2. Numerical information about selected latent variables. MI stands for mutual information, IC stands for information coverage. States of latent variables are denoted by $S0$ and $S1$. Other than MI and IC, the other decimal numbers are marginal probabilities of the states of the latent variables or the conditional probability distributions of the herb variables. See the discussion of $Y7$ to get the precise meanings of the terms and the numbers.

Y7	MI	IC	$S0$=.04	$S1$=.96
SFSD	.85	.85	.81	0
CP	.60	.98	.66	0

Y12	MI	IC	S0=.013	S1=.87
RC	1	1	.1	0
SCF	.16	1	.34	.02

Y11	MI	IC	S0=.95	S1=.06
MH	.71	.71	.05	1
HSBH	.70	.93	.02	.9
MCT	.46	.98	0	.56

Y20	MI	IC	S0=.79	S1=.21
SLAR	1	1	1	0
ULAR	.08	1	0	.12

Various numerical information about $Y7$ is given in Table 2. In terms of mutual information (MI), Y7 is the closest to SFSD and CP. The ratio of the MI between Y7 and SFSD over the MI between Y7 and all manifest variables is called the information coverage (IC) of SFSD with respect Y7. It is .85. Similarly, the information of the two variables SFSD and CP with respect Y7 is .98. Because of this, we can say that Y7 represents a partition of the prescriptions almost totally based on whether they use the two herbs SFSD and CP.

The partition consists of two classes $Y7$=$S0$ and $Y7$=$S1$. The former class has marginal probability 0.04. This indicates the class Y7=S0 consists of 4% the prescriptions. This class of prescriptions has high probabilities, 0.81 and 0.66 respectively, to prescribe the herbs SFSD and CP. On the other hand, the rest of the prescriptions do not use those two herbs at all.

These findings turn out to be very interesting. As a matter of fact, the combination of the two herbs SFSD and CP is a classical formula, called *Shi Xiao San*. It is mainly used to *promote blood circulation* and *remove blood stasis*. Therefore, the two herbs are often used for the treatment of DBLS patients accompanied with *blood stasis* syndrome.

Similar inspection reveals that states of other latent variables represent other interesting combinations of herbs: $Y11$=$S1$ represents the combination of *motherwort herb* and *medicinal cyathula toot*, which is for the DBLS patients with diseases, such as polycystic ovary syndrome and primary infertility; $Y12$=$S0$

represents the combination of *rhizoma corydalis and szechwan chinaberry fruit*, which is for the DBLS patients with symptoms of all kinds of pains; and so on.

Negative associations have also been uncovered. Prescriptions in the class $Y20{=}S0$ all use *stir-fry largehead atractylodes rhizome*, but never *uncooked largehead atractylodes rhizome*, while prescriptions in the class $Y20{=}S1$ have some probabilities of using the latter, but never former. This is consistent with reality. In practice, the two herbs are never used together. Other data mining methods such as association rules cannot find this kind of negative dependence.

5 Concluding Remarks

Prescription is a skill that TCM students take years to learn and master. By revealing patterns of herb combinations in the prescriptions by seasoned experts, our study can help the students to acquire the skill faster and better. They can also help pharmaceutical companies to decide what combination of Chinese herbs to test.

One future direction is to correlate the patterns that we found with symptoms. The results would be even more interesting to TCM researchers, practitioners, and students.

Acknowledgements. This work is partially supported by Program of Beijing Municipal S&T Commission, China (D08050703020803, D08050703020804), China NSFC project (90709006), National Key Technology R&D Program (2007BA110B06), and China 973 project (2011CB505101).

References

1. Anonymous: The Inner Canon of Emperor Huang. Chinese Medical Ancient Books Publishing House, Beijing (2003)
2. Tang, J.L., Liu, B.Y., Ma, K.W.: Traditional chinese medicine. Lancet 372 (2008)
3. Flaws, B., Sionneau, P.: The treatment of modern western medical diseases with Chinese medicine: a textbook and clinical manual, 2nd edn. Blue Poppy Press (2005)
4. Wang, L., Zhou, G.B., Liu, P., Song, J.H., Liang, Y., Yan, X.J., Xu, F., Wang, B.S., Mao, J.H., Shen, Z.X., Chen, S.J., Chen, Z.: Dissection of mechanisms of chinese medicinal formula realgar-indigo naturalis as an effective treatment for promyelocytic leukemia. PNAS 105 (2008)
5. Zhou, X., Chen, S., Liu, B., Zhang, R., Wang, Y., Li, P., Guo, Y., Zhang, H., Gao, Z., Yan, X.: Development of traditional chinese medicine clinical data warehouse for medical knowledge discovery and decision support. Artif. Intell. Med. 48(2-3) (2009)
6. Feng, Y., Wu, Z., Zhou, X., Zhou, Z., Fan, W.: Knowledge discovery in traditional chinese medicine: State of the art and perspectives. Artif. Intell. Med. 38(3) (2006)
7. Zhang, N.L., Yuan, S., Chen, T., Wang, Y.: Latent tree models and diagnosis in traditional chinese medicine. Artif. Intell. Med. 42 (2008)
8. Zhang, N.L., Yuan, S., Chen, T., Wang, Y.: Statistical validation of traditional chinese medicine theories. J. Altern. Complement Med. 14(5) (2008)

9. Jakulin, A., Bratko, I.: Testing the significance of attribute interactions. In: ICML 2004 (2004)
10. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics 6, 461–464 (1978)
11. Zhang, N.L.: Hierarchical latent class models for cluster analysis. Journal of Machine Learning Research 5, 697–723 (2004)
12. Zhang, R.: Clinical research on the syndrome structure and syndrome hierarchical differentiation of disharmony of liver and spleen syndrome. PhD dissertation, Guanganmen hospital, China academy of Chinese medical sciences (2008) (in chinese)