



Verifying TCM Syndrome Hypothesis Based on Improved Latent Tree Model

Nian Zhou^{1,2,3}, Lingshan Zhou⁴, Lili Peng^{1,2,3}, Bing Wang^{1,2,3(✉)},
Peng Chen⁵, and Jun Zhang⁶

¹ School of Electronics and Information Engineering,
Tongji University, Shanghai, China

zhounian@tongji.edu.cn, wangbing@ustc.edu

² The Advanced Research Institute of Intelligent Sensing Network,
Tongji University, Shanghai, China

³ The Key Laboratory of Embedded System and Service Computing,
Tongji University, Shanghai, China

⁴ Neurology Department, Jinzhou Medical University,
Shenyang, Liaoning, China
ls_zhou17@163.com

⁵ Institute of Physical Science and Information Technology,
Anhui University, Hefei 230601, China

⁶ School of Electronic Engineering and Automation,
Anhui University, Hefei 230601, China

Abstract. Traditional Chinese Medicine (TCM) is a significant channel for the prevention and treatment of Chinese diseases and is increasingly popular among non-Chinese people. However, it suffered serious credibility problems. The fundamental question is that TCM syndrome differentiation is it a totally subjective question or is it based on evidence? In recent years, a method called latent tree analysis (LTA) has been put forward. The main idea is, based on statistical principles for cluster analysis of the epidemiological survey symptoms data, to discover latent variables implicated in the data and compare them with TCM syndromes. However, LTA has its own limitations. It states that one manifest variable in the latent tree model (LTM) can only correspond to one latent variable. This is inconsistent with the theory of traditional Chinese medicine. Therefore, this paper proposed an improved LTA, based on the LTM obtained from the original LTA, adding arrows between symptoms and syndromes. The current analysis used the improved LTA to study a dataset of 37,624 patients with hepatopathy. The latent variables found here well match the latent factors of TCM, in addition, there are also some symptoms associated with multiple syndromes, it not only provides evidence for the validity of the relevant TCM hypothesis in the case of hepatopathy and helps to classify these patients into TCM syndromes, but also proved that the improved LTM has a higher degree fitting to the original data.

Keywords: Latent tree model · Syndrome differentiation · TCM

1 Introduction

Traditional Chinese Medicine is a significant channel for the prevention and treatment of Chinese diseases and is increasingly popular among non-Chinese people. However, it suffered serious credibility problems, especially in the West. One of the reasons is the lack of rigorous randomized trials to support the efficacy of TCM treatment [1]. Another equally important reason for this article is the lack of validation of TCM theory.

The diagnosis of TCM begins with a comprehensive observation of symptoms (including signs) using four diagnostic methods: examination, hearing, questioning and palpation. Based on the collected information, patients are classified as various types of TCM vocabulary collectively referred to as “Zheng” [2]. “Zheng” is usually translated as “Traditional Chinese Medicine Syndrome.” The process of classifying patients into various syndromes is called TCM syndrome differentiation.

TCM syndrome such as Yang deficiency, Yin deficiency is the hypothesis of traditional Chinese medicine to explain the occurrence and co-occurrence of signs and symptoms. For example, TCM believes that Yang and Yin are essential substances for the human body. They have the function of warming and nourishing the body [3]. Deficiency of yang can lead to cold performance, such as chilling. Therefore, patients with these symptoms are often classified as Yang class. Similarly, deficiency of Yin may lead to dry mouth, throat fever, fever of hands and feet, and other symptoms [4]. Therefore, patients with these symptoms are often divided into Yin deficiency and Yang deficiency class.

Western medicine divides patients into different categories according to disease types or subtypes and treats them accordingly. Instead, TCM classifies patients into different categories based on the type of symptoms and treats them accordingly. The difference between Chinese and Western medicine is that western medicine is treated by the type of disease, and TCM is treated by co-occurrence of symptoms [5–7].

Two fundamental questions about TCM syndromes are usually raised: (1) Do they correspond to real-world entities, or are they purely subjective? (2) TCM Syndrome Differentiation is it a totally subjective question or is it based on evidence? For more than half a century, researchers have been seeking answers to these questions through laboratory experiments [5, 6] However, the question remains unanswered [7–9].

Recently, a different method has been proposed [10, 11]. It distinguishes between two variables of traditional Chinese medicine. Clinically, symptoms such as “cold”, “dry mouth” and “throat dryness” can be directly observed and are therefore referred to as observation variables. On the other hand, the syndrome factors such as Yang deficiency and Yin deficiency cannot be directly observed and must be determined indirectly according to the symptoms. Therefore, they are latent factors.

Zhang et al. speculated that certain conceptions of syndromes such as Yang deficiency and Yin deficiency originated from the observed regularity of symptoms and co-occurrence in ancient times. They proposed a new TCM syndrome research method, in which the researchers collected data on the patient’s symptoms and ruled out the doctor’s diagnosis, and re-extracted, from the unlabeled data collected, the latent factors postulated in TCM. Since the purpose of this method is to provide objective

evidence for the diagnosis of TCM, the diagnostic results will not be collected in the first step. The second step is accomplished by using a new type of probability model called the Latent Tree Model, which was developed specifically for the study of TCM syndromes [10]. This method is called latent tree analysis (LTA).

They have tested the LTA method on the kidney deficiency data set. The latent variables they found were in good agreement with the relevant latent factors of TCM. (Note that in this article, the term “latent factor” refers to a factor that is not observed in TCM, and “latent variable” refers to a variable that is not observed in the statistical model.) This provides statistical validation of the relevant TCM hypothesis. However, this latent tree analysis has its own limitations. It states that one observed variable in the latent tree model can only correspond to one latent variable. This is inconsistent with the theory of traditional Chinese medicine. According to traditional Chinese medicine theory, a symptom may be caused by multiple Syndromes, that is, one symptom can be connected to multiple syndromes. Therefore, based on the model of LTM, by increasing the connection between symptoms and syndromes, the arrow-adding (LTM-AA) operation is performed to improve the fitting degree between the model and the data. Experiments have proved that the improved latent tree model has a higher degree fitting to the original data.

The current analysis used the improved LTA to study a dataset of 37,624 patients with hepatopathy. The latent variables found here well match the latent factors of TCM. This provides evidence for the validity of the relevant Chinese medicine hypothesis in the case of hepatopathy and helps to classify these patients into TCM syndromes.

2 Improved Latent Tree Model

2.1 Review Latent Tree Model

The latent tree model is the simplest latent structural model. They were previously called “hierarchical latent class models.” [12] The “latent tree model” is a tree-structured Bayesian network [13–15] in which the leaf node variables are observed and formed as “manifest variables,” whereas variables at internal nodes are latent and hence are called “latent variables.” All variables are assumed to be discrete. Arrows indicate direct probability dependence. In the model shown on the left side of Fig. 1, there is an arrow from variable Y1 to variable Y2. This means that Y2 is directly dependent on Y1. The dependence is characterized by a conditional distribution $P(Y2|Y1)$ which gives the Y2 distribution of each value of Y1.

As an example, consider high school students who need several subjects (such as math, science, literal, and history). Grades are influenced by latent factor analysis skills and literal skills, which in turn are affected by general intelligence. These relationships form a latent tree structure, as shown on the right side of Fig. 1.

The interest in this article is how to induce a latent tree model from the data. This can be divided into two sub-problems. First, which one is the best among all possible models? This is a matter of model selection. Zhang [12] empirically examined several criteria, namely, Bayesian Information Criterion (BIC) score [16], Akaike Information Criterion (AIC) score [17], Cheeseman-Stutz score [18], and holdout-likelihood [19].

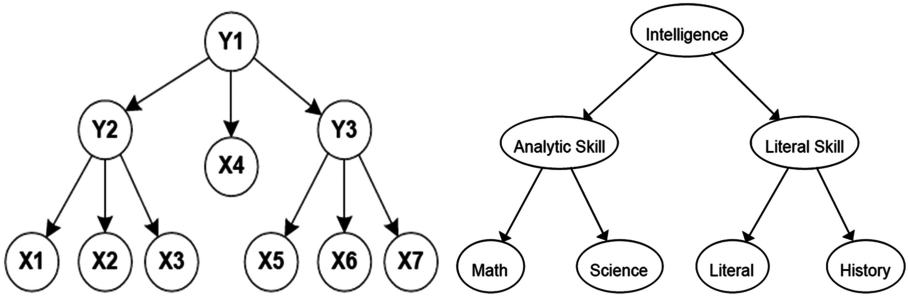


Fig. 1. Left: A latent tree model. The X_i 's are manifest variables and the Y_i 's are latent variables. Right: An example of latent tree model.

Bayesian Information Criteria (BIC) score turns out to be the most appropriate for the task. The BIC score for the latent tree model is given by formula (1):

$$\text{BIC}(m|D) = \log(D|m, \theta^*) - \frac{d(m)}{2} \log N \quad (1)$$

Where D is the data set, m is the structure of the model, θ^* is the maximum likelihood estimate of the model parameters, $d(m)$ is the number of free parameters, and N is the sample size. Please note that this definition of BIC score is used in the machine learning community rather than social science researchers who often use negatives.

The second sub-question is how to find the model with the highest BIC score in all possible model spaces. The first algorithm for this task was DHC [13]. It can handle data sets with only about six manifest variables. As for it, it is a concept test algorithm. The second algorithm is called SHC [20]. It can handle about a dozen manifest variables. It is a springboard for a more efficient EAST (Expansion Adjustment Simplification until Termination) [21] algorithm that can handle about 100 manifest variables. EAST has been tested on synthetic data and real-world datasets through market research [22] and from a social survey. It finds interesting latent structures in all situations.

2.2 The Arrow-Adding Algorithm

Based on the latent tree model obtained by the EAST algorithm, considering that one symptom is often associated with multiple syndromes in the TCM theory, this paper uses the arrow-adding (LTM-AA) to improve the original latent tree model. Improvement is made to enable one manifest variable to correspond to multiple latent variables, that is, one symptom corresponds to multiple syndromes. However, considering the complexity of the improved algorithm, this paper considers one syndrome can only add arrow once.

The LTM-AA algorithm is another search process under the guidance of the BIC score, adding one edge at each step. For example, for the latent tree model shown in left of Fig. 1, taking the latent variable Y_1 as an example, the AA algorithm is connected to X_1 , X_2 , X_3 , X_5 , X_6 , X_7 via Y_1 , respectively. Then there are six candidate models, comparing the BIC scores of the six candidate models and the initial model. and then

select the model with the largest BIC score as the initial model for the next step, the other latent variables Y2, Y3 search process is the same. In the end, the model shown in Fig. 2 may be obtained. This model is no longer a tree and is a more complex Bayesian network than the latent tree model.

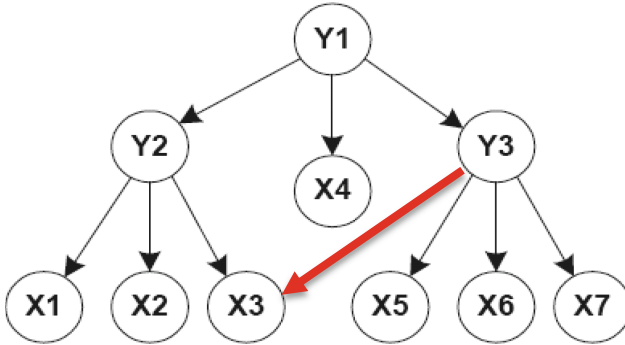


Fig. 2. The improved latent tree model which obtain by adding arrow algorithm, and the red arrow is the added edge. (Color figure online)

The LTM-AA algorithm is given as follows:

```

LTM-AA (m, θ*, D):
Repeat all latent variables:
  Repeat all manifest variables:
    (ml, θl*) ← Add arrow from latent variable to manifest variable.
    If BIC (m, θ*) ≤ BIC (ml, θl*),
      (m, θ*) ← (ml, θl*).
  Return (m, θ*).
  
```

3 Analysis of TCM Data Set

3.1 Data Collection

The dataset was collected in 2010–2014. It contains 37624 patient cases and 150 symptoms. These symptom variables were selected according to the Chinese National TCM Clinical Terminology Criteria [23] and some TCM diagnostic textbooks [24]. These variables are the most important factors for Chinese medicine doctors in determining whether a patient has hepatopathy. Therefore, we call this dataset hepatopathy data. Each symptom variable has four possible values, namely “no”, “light”, “medium” and “serious”. Operational criteria [25] are defined when collecting data to determine the consistency of the severity of the symptoms.

The data was from a Shanghai hospital and all subjects were inpatients or outpatients who had suffered from hepatopathy. Therefore, all the conclusions drawn are about this group.

3.2 Data Processing

Since the data sample is too large and the data itself is sparse, when the symptoms' frequency is less than 5%, the symptoms were removed from the data set, and there are remaining 135 symptoms. However, there are still too many symptoms, and the improved latent tree analysis can't handle such a big data, so we also removed some other symptoms which can be considered irrelevant symptoms. We removed them by measuring the similarity through Euclidean distance between the symptoms. After that, there are 55 symptoms contained.

In order to decrease the algorithm runtime, there are only 11854 patient cases chosen and the removed cases are which contains too few symptoms. Finally, 11854 patient cases and 55 symptoms included in further analysis.

3.3 The Result of LTA

We used the improved EAST algorithm to analyze hepatopathy data. The result of the analysis is a latent tree model that will be referred to as a hepatopathy model. The best model's BIC Score is -248106 and the structure of the model is shown in Fig. 3.

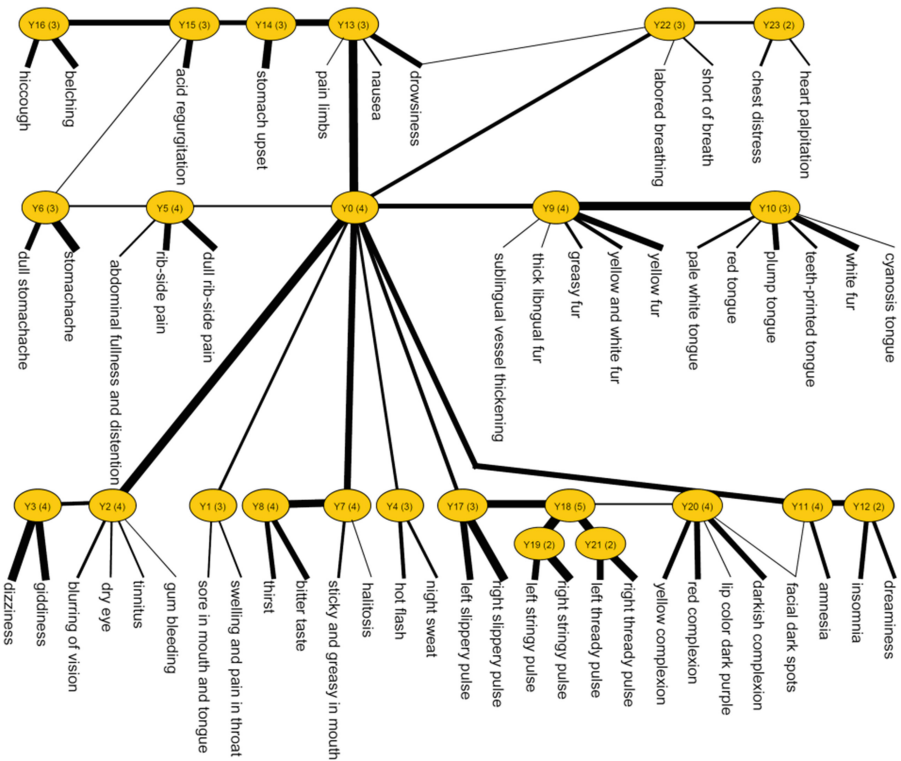


Fig. 3. The structure of the best model M found for hepatopathy data.

In the model, nodes marked with English phrases represent symptom variables. Each of them has four possible values that represent the degree of symptoms. The symptom variable comes from the data set. Nodes marked with a capital letter ‘Y’ and an integer index are latent variables. They are not from data sets. Instead, they are introduced in the data analysis process to explain the patterns in the data. Each latent variable has an integer next to it. This is the number of possible values for the latent variable.

The edges in the model represent probability dependencies. Each edge is characterized by a conditional probability distribution. The width of the edge indicates the strength of the correlation between the variables. For example, Y5 is strongly related to “rib-side pain”, related to “dull rib-side pain”, and weakly related to “abdominal fullness and distention.” In this article, we will focus on the links between variables and the advantages of these links. The conditional probability distribution contains quantitative information that can be used as a syndrome differentiation. We will discuss them in future work.

4 Discussions

4.1 Latent Variables as Evidence for TCM Hypothesis

In the latent tree model, the set of manifest variables that are directly connected to a particularly latent variable is called a sibling cluster. The siblings cluster together with the latent variables constitute a family. For example, two symptom variables under Y3, “dizziness” and “giddiness” form a sibling cluster. Together with Y3, they form a family and are said to be headed by Y3.

Why are some symptom variables grouped to form sibling clusters during latent tree analysis? Why introduce latent variables? Examination of the model (qualitative and quantitative information) shows that there are three conditions for this problem. First, some symptom variables are grouped into a sibling cluster because they tend to occur at the same time. In the family headed by Y8, one example is “thirst” and “bitter taste.” In this case, the latent variable is introduced to explain the co-occurrence of symptoms. Second, some symptom variables are grouped into a sibling cluster because they are mutually exclusive. One of the examples is “yellow complexion”, “red complexion” and “darkish complexion” headed by Y20. In this case, the latent variable is introduced to represent a partition of the patients based on those symptoms. The third case is a mix of the first two cases. An example is the family headed by Y10, where “red tongue” and “pale white tongue” are mutually exclusive, and they all occur together with “teeth-printed tongue”.

The latent variable in the first case is evidence of the validity of the TCM hypothesis. For example, traditional Chinese medicine believes that syndrome of deficiency of liver blood may lead to “dizziness” and “giddiness.” The implication of this hypothesis is that these two symptoms tend to co-occur in clinical practice. The introduction of Y3 in the latent tree analysis has confirmed that “dizziness” and “giddiness” really do tend to co-occur in the data. In other words, it has confirmed the meaning of this hypothesis. In this sense, it provides support for the assumption of traditional Chinese medicine.

4.2 Evidence for the Validity of TCM Hypothesis

Starting from the lower center of the model M, we noticed that ‘hot flash’ and ‘night sweat’ are grouped under Y4. This and related quantitative information suggest that these two symptoms tend to occur together in the data. On the other hand, the co-occurrence of these two symptoms is the hypothesis that fire excess from yin deficiency will lead to “hot flash” and “night sweat”. So, Y4 variable has confirmed this meaning and therefore proves the validity of this hypothesis.

The family headed by Y13 showed three symptoms of “pain limbs”, “nausea” and “drowsiness” tended to occur in the data at the same time. The evidence to support this hypothesis is that phlegm-heat blocking orifices may cause “pain limbs”, “nausea” and “drowsiness”.

The family headed by Y23 indicated that there was “chest distress” and “heart palpitation” co-occurrence in the data. The evidence supporting this hypothesis is that stagnant blockade of heart blood may lead to “chest distress” and “heart palpitation”. At the same time, family headed by Y22 show that in the data, together with “chest distress” and “heart palpitations”, there are also symptoms of “labored breathing”, “short of breath” and “drowsiness” occurred together. This is evidence supporting the hypothesis that chest disorder of qi movement may lead to “chest distress”, “heart palpitation”, “labored breathing”, “short of breath” and “drowsiness”.

In addition, we found that the symptoms of “drowsiness” were linked to Y13 and Y22 at the same time, and that the interpretation of Y13 and Y22 could both correspond to the syndromes in TCM theory. Y22 and Y13 were related to each other, and it was consistent with the hypothesis that phlegm-heat blocking orifices is relevant to chest disorder of qi movement. This not only verifies the correctness of the latent tree model, but also shows that the improved latent tree model has higher degree of fitting to the original data.

Family headed by Y12 showed two symptoms: “insomnia” and “dreaminess” tend to occur in the data at the same time. The evidence supporting this hypothesis is that yin deficiency can lead to “insomnia” and “dreaminess”. The family headed by Y5 indicated that there is “stomachache”, “dull stomachache”, “abdominal fullness and distention”, “rib-side pain” and “dull rib-side pain” co-occurrence in the data. The theory of syndrome differentiation in traditional Chinese medicine generally considered that: (1) “stomachache” and “dull stomachache” suggest that the lesion is located on the stomach, which is a manifestation of stomach qi stagnation. There are many causes of stomach qi stagnation, such as phlegm, drinking, hydrosphere, moisture, retained food, cold, heat and other evil stagnation in the stomach, or stomach yang deficiency, can lead to stomach qi stagnation and painful. (2) abdominal fullness and distention prompted lesions in the spleen, but also in the large intestine, liver, etc. it often due to a variety of deficiency, such as yang, qi deficiency and real qi stagnation, coagulation cold, congestion and other factors cause local blood block. (3) “rib-side pain” and “dull rib-side pain” suggest that the lesions are mostly in the liver and gallbladder, are often caused by pathogenic factors such as qi stagnation, hot and humid condition, heat pathogen, etc. which hinder the liver and gallbladder vent. The evidence supporting this hypothesis is that spleen and stomach qi stagnation and liver qi stagnation may lead to the above five symptoms.

The other sibling clusters in the model are also clearly meaningful. The variables under Y9 are mostly about the color of fur; the variables under Y10 are mostly about the situation of the tongue; And the variables of Y20 represents a partition of “color of complexion”.

5 Conclusion

Based on the latent tree model obtained by EAST algorithm, an arrow-adding (LTM-AA) algorithm is used to achieve a single manifest variable that can correspond to multiple latent variables, that is, one symptom corresponds to multiple syndromes in TCM theory.

An improved LTM was performed on the symptom data of 11854 patients with hepatopathy. This article introduces the model obtained and explains how to understand and appreciate the qualitative aspects of the model. In particular, this report discusses how and in what sense data analysis provides evidence for confirming the TCM syndrome hypothesis. All about this evidence is determined through systematic inspection of the model.

This analysis shows that according to the data, a symptom corresponds to multiple syndromes with their realistic basis. For example, for symptoms of drowsiness, before the arrow-adding algorithm is performed, the drowsiness is only associated with the latent variable Y13, but by the arrow-adding algorithm, it can find that it is also related to the latent variable Y22, in addition, for the latent variables Y13 and Y22, we can find the corresponding syndrome in the TCM. This not only verifies the correctness of the traditional Chinese medicine theory for chest disorder of qi movement, phlegm-heat blocking orifices, but also reflects the improvement of the latent tree model is more fit to the original data. In the same way, this work has provided statistical validation to TCM hypothesis about yin deficiency, fire excess from yin deficiency, chest disorder of qi movement, phlegm-heat blocking orifices, stagnant blockade of heart blood, stomach qi stagnation and liver qi stagnation.

References

1. Normile, D.: The new face of traditional Chinese medicine. *Science* **299**(5604), 188–190 (2003)
2. World Health Organization: WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region. WHO Regional Office for the Western Pacific, Manila (2007)
3. Wu, D.X., Li, D.X., Yan, S.Y.: *Fundamental Theories of Traditional Chinese Medicine*. Science and Technology Press, Shanghai (1994)
4. Zhu, B., Wang, H.: *Diagnostics of Traditional Chinese Medicine*. Singing Dragon, London (2011)
5. Wang, H., Xu, Y.: *The Current State and Future of Basic Theoretical Research on Traditional Chinese Medicine*. Military Medical Sciences Press, Beijing (1999)
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), 219–236 (2006)

7. Liang, M., Liu, J., Hong, Z., Xu, Y.: *Perplexity of TCM Syndrome Research and Countermeasures*. People's Health Press, Beijing (1998)
8. Wang, B., Shen, H., Fang, A., D.-s., H., Jiang, C., Zhang, J., et al.: A regression model for calculating the second dimension retention index in comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry. *J. Chromatogr. A* **1451**, 127–134 (2016)
9. Wang, B., Chen, P., Wang, P., Zhao, G., Zhang, X.: Radial basis function neural network ensemble for predicting protein-protein interaction sites in heterocomplexes. *Protein Pept. Lett.* **17**(9), 1111–1116 (2010)
10. Zhang, N.L., Yuan, S., Chen, T., Wang, Y.: Latent tree models and diagnosis in traditional Chinese medicine. *Artif. Intell. Med.* **42**(3), 229–245 (2008)
11. Zhang, N.L., Yuan, S., Chen, T., Wang, Y.: Statistical validation of traditional Chinese medicine theories. *J. Altern. Complement. Med.* **14**(5), 583–587 (2008)
12. Zhang, N.L.: Hierarchical latent class models for cluster analysis. *J. Mach. Learn. Res.* **5**(6), 697–723 (2004)
13. Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier, New York (2014)
14. Chen, P., Hu, S., Zhang, J., Gao, X., Li, J., Xia, J., et al.: A sequence-based dynamic ensemble learning system for protein ligand-binding site prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB)* **13**(5), 901–912 (2016)
15. Xia, S., Chen, P., Zhang, J., Li, X., Wang, B.: Utilization of rotation-invariant uniform LBP histogram distribution and statistics of connected regions in automatic image annotation based on multi-label learning. *Neurocomputing* **228**, 11–18 (2017)
16. Schwarz, G.: Estimating the dimension of a model. *Ann. Stat.* **6**(2), 461–464 (1978)
17. Akaike, H.: A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**(6), 716–723 (1974)
18. ACSR: Bayesian classification (autoclass): theory and results (1996)
19. Cowell, R.G., Dawid, P., Lauritzen, S.L., Spiegelhalter, D.J.: *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Springer, New York (2006). <https://doi.org/10.1007/b97670>
20. Zhang, N.L., Kocka, T., (eds.): Efficient learning of hierarchical latent class models. In: 16th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2004. IEEE (2004)
21. Chen, T., Zhang, N.L., Liu, T., Poon, K.M., Wang, Y.: Model-based multidimensional clustering of categorical data. *Artif. Intell.* **176**(1), 2246–2269 (2012)
22. Zhang, N.L., Yi, W., Tao, C.: Discovery of latent structures: experience with the CoIL challenge 2000 data set. *J. Syst. Sci. Complex.* **21**(2), 172–183 (2008)
23. Supervision CSBoT: National Standards on Clinic Terminology of Traditional Chinese Medicinal Diagnosis and Treatment—Syndromes. China Standards Press, Beijing (1997)
24. Yang, W.M.F., Jiang, Y.: *Diagnostics of Traditional Chinese Medicine*. Academy Press, Beijing (1998)
25. Yan, S.L., Zhang, L.W., Wang, M.H., Yuan, S.H.: Operational standards for determining the severity levels of kidney deficiency symptoms. *J. Chengdu Univ. Chin. Med.* **24**(1), 56–59 (2001)