Journal homepage:
www.jcimjournal.com/jim
www.elsevier.com/locate/issn/20954964
Available also online at www.sciencedirect.com.

• Methodology

# A data–driven method for syndrome type identification and classification in traditional Chinese medicine

**Nevin Lianwen Zhang**[1], **Chen Fu**[2], **Teng Fei Liu**[1], **Bao-xin Chen**[2], **Kin Man Poon**[3], **Pei Xian Chen**[1], **Yun-ling Zhang**[2]

1. Department of Computer Science and Engineering, the Hong Kong University of Science and Technology, Hong Kong, China
2. Department of Neurology, Dongfang Hospital, Beijing University of Chinese Medicine, Beijing 100078, China
3. Department of Mathematics and Information Technology, the Education University of Hong Kong, Hong Kong, China

## ABSTRACT

The efficacy of traditional Chinese medicine (TCM) treatments for Western medicine (WM) diseases relies heavily on the proper classification of patients into TCM syndrome types. The authors developed a data-driven method for solving the classification problem, where syndrome types were identified and quantified based on statistical patterns detected in unlabeled symptom survey data. The new method is a generalization of latent class analysis (LCA), which has been widely applied in WM research to solve a similar problem, i.e., to identify subtypes of a patient population in the absence of a gold standard. A well-known weakness of LCA is that it makes an unrealistically strong independence assumption. The authors relaxed the assumption by first detecting symptom co-occurrence patterns from survey data and used those statistical patterns instead of the symptoms as features for LCA. This new method consists of six steps: data collection, symptom co-occurrence pattern discovery, statistical pattern interpretation, syndrome identification, syndrome type identification and syndrome type classification. A software package called Lantern has been developed to support the application of the method. The method was illustrated using a data set on vascular mild cognitive impairment.

**Keywords:** medicine, Chinese traditional; syndrome; syndrome classification; latent tree analysis; symptom co-occurrence patterns; patient clustering; stand syndrome differentiation

**Citation:** Zhang NL, Fu C, Liu TF, Chen BX, Poon KM, Chen PX, Zhang YL. A data-driven method for syndrome type identification and classification in traditional Chinese medicine. *J Integr Med*. 2017; 15(2): 110–123.

## 1 Introduction

Traditional Chinese medicine (TCM) has been increasingly used in healthcare in China and around the world as a complementary or alternative method to Western medicine (WM).[1] A common practice is to divide the patients with a WM disease into several TCM syndrome types based on symptoms and signs (both referred as symptoms henceforth for simplicity), and to apply different TCM treatments to patients of different

http://dx.doi.org/10.1016/S2095-4964(17)60328-5
**Received** August 30, 2016; **accepted** November 7, 2016.
**Correspondence:** Prof. Nevin Lianwen Zhang; E-mail: lzhang@cse.ust.hk. Prof. Yun-ling Zhang; E-mail: yunlingzhang2004@163.com

types. The efficacy of TCM treatments depends heavily on whether the classification is done properly.[2,3]

The problem of TCM syndrome classification of a WM disease consists of four subproblems:[4–7] (1) What TCM syndrome types exist among the patients with the disease? (2) What is the prevalence of each syndrome type? (3) What are the characteristics of each syndrome type in terms of symptom occurrence probabilities? (4) How do we determine to the syndrome type(s) of a patient, based on symptoms?

The syndrome classification problem is of fundamental importance to TCM research and clinical practice.[2,3] As will be seen in Section 9, this problem has so far not been satisfactorily solved. In this paper we present a data-driven method for its solution. The idea is to: (1) conduct a cross-sectional survey of the patients with the WM disease and collect information about symptoms of interest to TCM; (2) perform cluster analysis on the data and divide the patients into clusters based on symptom occurrence patterns; (3) match the patient's clusters with TCM syndrome types; (4) use the statistical characteristics of the patient's clusters to quantify the TCM syndrome types and to establish classification rules.

We will first explain the data analysis methods that this paper relies upon in Section 2. Then we will present our method for solving TCM syndrome classification in Sections 3 through 8. Related works and limitations will be discussed in Section 9 and conclusions drawn in Section 10. A data set on vascular mild cognitive impairment (VMCI)[8] will be used for illustration throughout the paper.

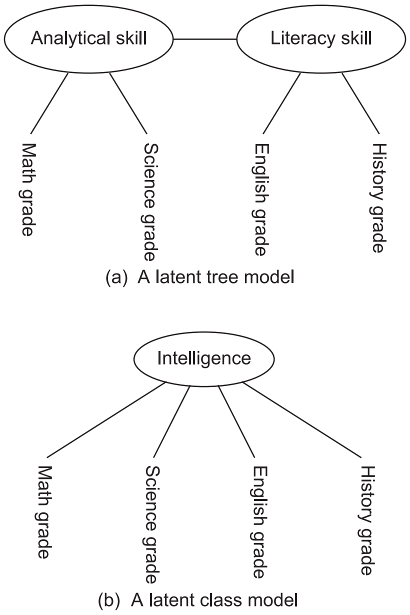
## 2 Technical background

This paper builds upon two data analysis methods, namely latent class analysis (LCA) and latent tree analysis (LTA). They are based on probabilistic models that describe relationships among categorical variables. Some of the variables are observed, while the others are latent, that is, unobserved. In this section, we explain LCA and LTA in layman's term so that medical researchers without a strong background in statistics and machine learning can understand the key ideas.

### 2.1 Latent tree models and latent class models

The models that we use are called latent tree models (LTMs). An LTM describes the relationship among a set of variables at two levels. At the qualitative level, it is an undirected tree where the observed variables are located at the leaf nodes, whereas the latent variables are located at the internal nodes. At the quantitative level, it describes the relationship between each pair of neighboring variables using a conditional probability distribution.

Figure 1(a) shows an example LTM taken from Xu et al.[9] Qualitatively, it asserts that a student's math grade (MG) and science grade (SG) are influenced by his analytical skill (AS); his English grade (EG) and history grade (HG) are influenced by his literacy skill (LS) and the two skills are correlated. Here, the grades are observed variables, while the skills are latent variables.



(a) A latent tree model

(b) A latent class model

| P(MG\|AS) | MG = low | MG = high |
|---|---|---|
| AS = low | 0.8 | 0.2 |
| AS = high | 0.2 | 0.8 |

| P(AS) | AS = low | AS = high |
|---|---|---|
| | 0.7 | 0.3 |

| P(LS\|AS) | LS = low | LS = high |
|---|---|---|
| AS = low | 0.6 | 0.4 |
| AS = high | 0.4 | 0.6 |

**Figure 1** The concept of latent tree model and latent class model

The subfigure (a) and the tables illustrate the concept of latent tree models using an example that involves two latent variables (the skill variables) and four observed variables (the grade variables). The tables show some of the probability parameters for the latent tree model. The subfigure (b) illustrates the concept of latent class models where intelligence is the only latent variable.

AS: analytical skill; LS: literacy skill; MG: math grade.