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

Nevin L. Zhang¹, Runsun Zhang², and Tao Chen³

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
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.