Journal of Biomedical Informatics 58 (2015) 260–267

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Discovering treatment pattern in Traditional Chinese Medicine clinical cases by exploiting supervised topic model and domain knowledge

Liang Yao, Yin Zhang *, Baogang Wei, Wei Wang, Yuejiao Zhang, Xiaolin Ren, Yali Bian

College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 31 March 2015
Revised 7 September 2015
Accepted 22 October 2015
Available online 30 October 2015

Keywords:
Traditional Chinese Medicine
Clinical cases
Treatment pattern discovery
Topic model
TCM domain knowledge

ABSTRACT

In Traditional Chinese Medicine (TCM), the prescription is the crystallization of clinical experience of doctors, which is the main way to cure diseases in China for thousands of years. Clinical cases, on the other hand, describe how doctors diagnose and prescribe. In this paper, we propose a framework which mines treatment patterns in TCM clinical cases by exploiting supervised topic model and TCM domain knowledge. The framework can reflect principle rules in TCM and improve function prediction of a new prescription. We evaluate our method on 3090 real world TCM clinical cases. The experiment validates the effectiveness of our method.

1. Introduction

As a complete medical knowledge system other than orthodox medicine, Traditional Chinese Medicine (TCM) plays an indispensable role in health care for Chinese people for several thousand years [1]. In TCM, a prescription is a set of medicines. In the long Chinese history, a large number of prescriptions have been invented to cure diseases [2]. Clinical cases, on the other hand, describe how doctors diagnose and prescribe. These cases recorded in ancient textbooks or hospitals are main TCM knowledge sources for generation of appropriate clinical hypotheses [3].

In the clinical practice of TCM, the rules of "Li-Fa-Fang-Yao" [4,5] are of critical importance. Li-Fa-Fang-Yao, which means principles, methods, prescriptions and Chinese herbal medicines, indicates basic steps of diagnosis and treatment: determine the cause and mechanism of diseases according to medical theories and principles, then decide treatment principles and methods, and finally select a prescription as well as proper Chinese herbal medicines.

In this paper, we propose a framework which can mine treatment patterns automatically from TCM clinical cases. Firstly, corresponding to "Li" (principles) and "Fa" (methods), we map symptoms to syndromes and determine treatment methods with TCM domain ontology. Then we mine medicine usage patterns in prescriptions via probabilistic topic model, which is corresponding to "Fang" (prescriptions) and "Yao" (medicines). Finally, these patterns could be used to improve the function prediction of a prescription in a new clinical case.

The rest of the article is organized as follows. Section 2 summarizes some related studies. Section 3 describes steps for mining treatment patterns. Section 4 carefully presents our experimental results and the result analysis. Finally, some conclusions and future works are provided in Section 5.

2. Background

Knowledge discovering and data mining have become very popular in health care and biomedicine [6]. Compared with clinical data mining research in modern biomedicine, TCM clinical data mining only becomes a hot topic in recent years. The related work of TCM knowledge discovery has been reviewed by Feng et al. [7] and Lukman et al. [8]. Zhou et al. [3] has reviewed text mining studies in TCM.

Specifically, Zhou et al. [9] introduced a clinical data warehouse system, which incorporates structured electronic medical record data for medical knowledge discovery and TCM clinical decision support. To illuminate the statistical foundation and objective diagnosis standards of syndrome differentiation, Zhang et al. [10] proposed the latent tree model to learn diagnosis structures from

* Corresponding author.
E-mail addresses: yaoliang@zju.edu.cn (L. Yao), yinzh@zju.edu.cn (Y. Zhang), wbg@zju.edu.cn (B. Wei), stephenwangw@163.com (W. Wang), yjzhangshi@163.com (Y. Zhang), xiaolinr@126.com (X. Ren), bianyali@zju.edu.cn (Y. Bian).

http://dx.doi.org/10.1016/j.jbi.2015.10.012
1532-0464/

the TCM clinical data sets with symptom variables in an unsupervised way, the latent tree model is similar to the topic model we employ. Based on the principle of co-occurrence, TCM concept networks are built in [5,11].

Some previous studies have devoted to studying the component law of medicines in prescriptions. For example, to analyze the component law of Chinese medicines and develop new prescriptions for gout, core combinations of herbs and new prescriptions were analyzed by using modified mutual information, complex system entropy cluster and unsupervised hierarchical clustering respectively in [12]. Yang et al. [13] presented a hierarchical clustering algorithm to discover novel prescriptions in TCM from formulae data and developed prescriptions from pharmacology by regression methods.

Probabilistic topic models such as Latent Dirichlet Allocation (LDA) [14] are widely-used statistical models that could find latent topics in documents. It could also be used in health care and biomedicine. For instance, Huang et al. [15] introduced an approach for discovering latent clinical process patterns from careflow logs. The main idea is to collect careflow logs and then estimate latent patterns for the collected logs based on LDA. Van Esbroeck et al. [16] explored the application of topic models to heart rate time series to identify functional sets of heart rate sequences and to concisely describe patients using task-independent features for various cardiovascular outcomes. The closest works to ours are [17,18]. Zhang et al. [17] proposed a multi-relational topic model by extending the author-topic model to incorporate multiple relationships between symptoms, herbal prescriptions and diagnoses. The proposed model is useful to extract underlying common symptom groups of one specific general disease classification, like diabetes, stroke and heart disease. Jiang et al. [18] applied LinkedLDA to symptom-herb topic detection. Our work is different from theirs because we focus more on syndromes and treatment methods, which is more consistent with the rules of "Li-Fa-Fang-Yao".

3. Method

The overall framework is shown in Fig. 1, with details presented in following subsections.

1. Given a collection of TCM clinical cases, we identify symptoms and medicines, label syndromes on each case, and decide treatment methods with TCM domain ontology. This step is corresponding to "Li" (principles) and "Fa" (methods).
2. After syndrome labeling and treatment methods determining, we mine treatment patterns, especially medicine usage patterns in prescriptions, via supervised topic model. This step is corresponding to "Fang" (prescriptions) and "Yao" (medicines).
3. We use medicines and prescription topics as features to predict treatment methods within the prescription. Syndromes cured by a new prescription can be inferred from the treatment methods.

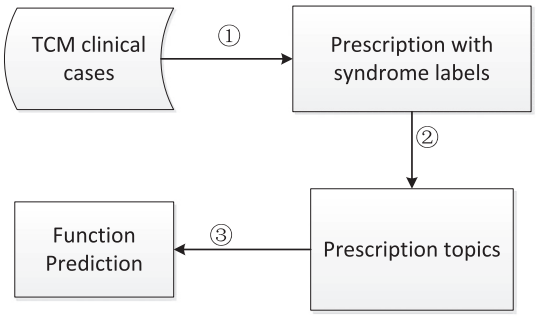

Fig. 1. Treatment pattern discovering framework.

3.1. Syndrome labeling and treatment methods determining

This step is corresponding to "Li" and "Fa", which means determining syndromes by observing a patient's symptoms and selecting treatment methods to cure syndromes. As a syndrome can be inferred from symptoms, there might be some connections between networks of symptoms and syndromes.

We label syndromes on clinical cases by mapping symptoms in a clinical case to syndromes in TCM domain ontology, like the Concept Labeling method [19] which builds text classifiers. In our TCM ontology based on Traditional Chinese Medical Subject Headings (TCM MeSH) [20],[1] a syndrome is a category, which contains a list of symptoms.[2] There are some symptoms or syndromes mentioned in a clinical case. We map these symptoms or syndromes to syndrome categories, categories which have been mapped more than a threshold value will be syndromes of a patient in the clinical case. Formally, the mapping is defined as follows.

An ontology $O = (V, E, \psi_{ont})$ is defined as a triplet: (1) a set of concepts V, (2) a graph of directed edges E that captures relationships between concepts, i.e., an edge $(v_1, v_2) \in E$ indicates that v_2 is a sub-concept of v_1, and (3) features of each concept V, ψ_{ont}. With respect to our TCM domain ontology based on *TCM MeSH*, concepts are syndromes or symptoms, the edge means v_2 is a symptom of syndrome v_1, or symptom v_1 and symptom v_2 co-exist. For instance, "constipation" is a symptom of "interior heat syndrome", symptom "constipation" and "thirsty" co-exist in "interior heat syndrome". The probability of a clinical case c labeled by syndrome s is

$$P(s|c) = \sum_{v \in V} P(s, v|c) = \sum_{v \in V} P(v|c)P(s|v) \quad (1)$$

where $P(s, v|c)$ is the probability of concept v and syndrome s given clinical case c, when the probability of v is determined, the probability of s can be determined by syndrome categories, $P(v|c)$ is clinical case to concept (symptom/syndrome) distribution, which can be calculated through text matching, when v is a symptom or syndrome, $P(v|c)$ can be treated as number of times v occurs in clinical case c, and $P(s|v)$ is concept to syndrome distribution, which can be calculated with the structure of TCM ontology, specifically, when v is a symptom, $P(s|v)$ could be interpreted as the number of times syndrome s contains v; when v is a syndrome and v is the same to s, $P(s|v)$ equals to 1, otherwise 0. After determining syndromes, treatment methods are easily determined by TCM domain knowledge. For example, when the syndrome of blood stasis is approved, then the treatment method of "activating blood and resolving stasis" is determined.

3.2. Topic discovering

In this section, we propose a supervised topic model based method to mine the herbal medicine usage patterns in TCM prescriptions. We introduce some notations and terminologies at first. Then we present our approach in detail.

3.2.1. Notation and terminology

Let M be the set of herbal medicines, a prescription p in a clinical case is a set of herbal medicines, i.e., $p = (m_1, m_2, \ldots, m_{N_p})$,

[1] Available at http://zcy.ckcest.cn/tcm/dic/home.

[2] The detailed syndrome categories is available at https://github.com/yao8839836/formulae/blob/master/formulae/src/file/syndrom_symptom.txt.

where $m_i \in M (1 \leqslant i \leqslant |M|)$ is a particular herbal medicine which can be filtered by text matching. Each prescription is associated with a list of binary treatment method label (topic) presence/absence indicators $\Lambda^p = (l_1, \ldots, l_k)$ and each $l_k \in \{0, 1\}$. Here N_p is the number of medicines in p, $|M|$ is M's size and K is the total number of unique treatment method labels.

With respect to our topic model based method, medicines are "words" in the model. A prescription prescribed in a clinical case is a bag of medicines, and we treat it as a "document" in our model. Treatment methods are "topics" of the "document". And a "corpus" is a collection of these "documents".

3.2.2. Generative process

After determining treatment methods, doctors need to select herbal medicines to form a prescription. Usually, a group of medicines is selected to cure a particular syndrome and is corresponding to a treatment method, so, to cure a patient's syndromes, several groups of medicines are selected. In this way, a prescription is a mixture of "topics", each "topic" is corresponding to a treatment method. The generative process is shown in Fig. 2.

This process is analogous to the generative process of probabilistic topic model. Topic models, like Latent Dirichlet Allocation (LDA) [14], model each document as a mixture of underlying topics and generates each word from one topic. Labeled LDA [21], as an extension to LDA, associates each label with one topic in direct correspondence and incorporates supervision by simply constraining the topic model to use only those topics that correspond to a document's label set.

In this study, we employ Labeled LDA to mine medicine usage patterns, and we set the number of topics in Labeled LDA to be the number of unique treatment methods K in the clinical cases. The generative process is shown as follows:

1. Draw K multinomials ϕ_s over medicines from a Dirichlet distribution with prior parameter β, one for each treatment method s;
2. For each prescription p:
   (a) For each method $s \in \{1, \ldots, K\}$, generate treatment method label (topic) presence/absence indicators from a Bernoulli distribution, i.e., $\Lambda_s^p \in \{0, 1\} \sim Bernoulli(\cdot|\Phi_s)$, Φ_s is the label prior for s;
   (b) Generate Dirichlet prior vector $\vec{\alpha}^{(p)}$ given label presence/absence indicators $\Lambda^{(p)}$ and predefined Dirichlet priors $\vec{\alpha}$, i.e., $\vec{\alpha}^{(p)} = \Lambda^{(p)} \times \vec{\alpha}$;
   (c) Generate treatment method mixture θ_p from Dirichlet distribution $Dir(\vec{\alpha}^{(p)})$, i.e., $\theta_p \sim Dir(\vec{\alpha}^{(p)})$.
   (d) For each i in $\{1, \ldots, N_p\}$
      (i) Generate a treatment method s from multinomial distribution $Mult(\theta_p)$, i.e., $s \sim Mult(\theta_p)$;
      (ii) Generate a medicine m_i from multinomial distribution $Mult(\phi_s)$, i.e., $\sim Mult(\phi_s)$;

In this process, treatment method labels associated with a prescription are used to project the Dirichlet prior vector $\vec{\alpha} = (\alpha_1, \ldots, \alpha_K)$ to a lower dimension $\vec{\alpha}^{(p)}$. The dimension of the projected vector corresponds to topics represented by treatment method labels of the prescription. For instance, suppose there are $K = 5$ treatment methods and a prescription p has treatment method labels given by $\Lambda^p = \{0, 1, 0, 1, 0\}$ which implies p's treatment method label set $\lambda^p = \{2, 4\}$, then treatment method mixture θ_p is drawn from a Dirichlet distribution with prior vector $\vec{\alpha}^{(p)} = L^{(p)} \times \vec{\alpha} = (\alpha_2, \alpha_4)^T$.

The model is same as traditional LDA, except the constraint that the topic prior $\vec{\alpha}^{(p)}$ is now restricted to the set of labeled treatment methods λ^p.

The exact inference for Labeled LDA is intractable, several approximate schemes have been developed to infer the model, in this study, we use collapsed Gibbs sampling [22] to estimate the probability a treatment method k assigned to the ith medicine m_i in a prescription p. In each iteration of Gibbs Sampling, the probability is given by:

$$P(z_{pi} = k|\vec{z}_{-i}) \propto \frac{n_{pk} + \alpha_k}{n_p + \vec{\alpha}^T \vec{1}} \times \frac{n_{km_i} + \beta}{n_k + |M|\beta} \tag{2}$$

where z_{pi} is the treatment method assignment of medicine m_i in prescription p, $\vec{z}_{-i}$ is all medicines' treatment method assignment excluding current m_i, n_{km_i} is the number of times medicine m_i is assigned to treatment method k, n_k is the total number of medicines assigned to treatment method k, n_{pk} is the number of times medicines in p assigned to treatment method k and n_p is the number of medicines in p.

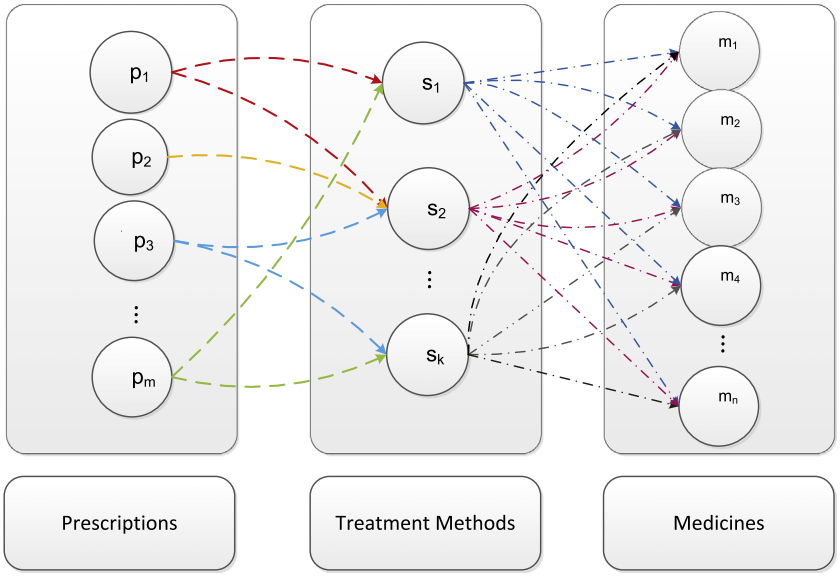

Fig. 2. The generative process of prescriptions in clinical cases.

三、通腑下瘀、涤化痰浊法治愈中风一例
(Treating stroke by dredging fu-organs, removing blood stasis and phlegm turbidity)

赵某，男，40岁。
(Zhao, Male, 40 years old.)

主诉及病史(Chief complaint and history)：卒然昏仆，不省人事(unconsciousness)，肢痉，遗尿(Enuresis)，痰声辘辘(heavy phlegm)，大便秘结(constipation)不行。
(Chief complaint and history: Suddenly fainted, unconsciousness, limbs convulsion, enuresis, and constipation.)

诊查 (Diagnosis and examination)：血压高至200/130mmHg，脉息滑数。撬齿视苔黄腻。询知平素嗜酒、吸烟。
(Diagnosis and examination: blood pressure 200/130mmHg, slippery and rapid pulse. Yellowish & greasy tongue fur. The patient usually drinks alcohol and smokes.)

辨证(Differentiation)：其证显系痰热肝火随气上逆，激犯清空，血络阻滞(blood block)、瘀(stasis)闭清窍所致，为中风(wind stroke)闭实之重症。正如《内经》中所谓："血之与气，并走于上，则为大厥，厥则暴死；气复反则复生，不反则死。"
(Differentiation: phlegm-heat and liver fire upward reverse with qi, and invade clear orifices, blood blocked, clear orifices confused by stasis, it's severe wind-stroke block syndrome. As "Inner Canon of Yellow Emperor" said: blood and qi go upward, it's major syncope, the syncope leads to sudden death; If qi goes back, people would revive, otherwise die.)

治法 (Treatment Method)：急当泄其上逆之气血痰火，乃亟投桃仁承气合温胆汤以通腑下瘀，涤化痰浊。
(We should discharge the upward qi, blood and phlegm-fire immediately, and then used Tao Ren Chen Qi He Wen Dan Decoction to dredge fu-organs, remove blood stasis and phlegm turbidity.)

处方 (Prescription)：生大黄(Rhubarb) 10g 芒硝(Sodium Sulfate) 10g 桃仁(Peach Seed) 10g 竹沥 (Liquor Phyllostachys) 半夏 (Pinellia Tuber) 10g 陈皮 (Dried Tangerine Peel) 6g 茯苓 (Poria) 12g 甘草 (Liquorice Root) 3g 枳实 (Immature Orange Fruit) 10g 石菖蒲 (Grassleaf Sweetflag Rhizome) 10g 钩藤 (Gambir plant Nod) 12g（后下）炙远志 (Milkwort Root) 6g 竹沥水20ml冲服

药后大便排出多量粪块，神志转清，痉定。唯右侧肢体偏瘫 (Hemiplegia)，续予涤痰化浊之剂，用指迷茯 苓丸及和营通络之品，调治半月，逐渐恢复，行步自如。越8年后方殁。
(After taking medicines, the patient stooled volume turd and became conscious, only right limb were paralyzed. We continued providing prescriptions which remove blood stasis and phlegm turbidity, and used Zhi Mi Fu Ling pills and medicines which harmonize the nutrient and dredge collaterals. After nursing the patient's health for half month, he recovered gradually and walked freely. The patient died 8 years later.)

[按语] (Note) 通腑祛瘀、清化痰热是治疗中风痹多种疗法之一，临证当以具有瘀、热、痰等邪实的 病理表现为据，运用桃仁承气汤合温胆汤，下其瘀热、泻其痰火，往往能获气平血降、痰消火清之 效，从而使病情转危为安，迅速向愈。此外，桃仁承气汤合温胆汤治疗中风，不但适用于瘀、热、痰 的闭证，对闭象解除后的后遗症，及中经络症，有瘀、热、痰等病理表现者，亦同样可以取用。
(Note: Dredging fu-organs, removing blood stasis and phlegm turbidity are treatment methods for stroke. Syndromes should have evil domination symptoms like stasis, heat and phlegm. Using Tao Ren Chen Qi He Wen Dan Decoction to remove the stasis, heat, to discharge phlegm-fire, could result in a good effect and cure the disease. Additionally, Tao Ren Chen Qi He Wen Dan Decoction is not only suitable for wind-stroke block syndrome, but also suitable for sequelas, collateral symptoms and other syndromes which has symptoms like heat, stasis and phlegm.)

Fig. 3. An example clinical case.

After Gibbs Sampling iterations, we estimate the treatment method-medicine multinomials ϕ and the prescription-treatment method mixture weights θ from the final $\vec{z}$ sample, using the means of their posteriors given by

$$\phi_k(m) \propto n_{km} + \beta \tag{3}$$

$$\theta_p(k) \propto n_{pk} + \alpha_k \tag{4}$$

The treatment method-medicine multinomials ϕ_k for each treatment method k are our learned "topics" and can be represented by medicines with high probability; each prescription-treatment method multinomial θ_p represents the prevalence of treatment methods within prescription p.

3.3. Function prediction

After topic discovering, we can infer the prevalence of treatment methods (topics) within a new prescription by using Bayes rule:

$$P(k|\vec{m}) \propto \prod_{m_i \in \vec{m}} P(m_i|k)P(k) = \prod_{m_i \in \vec{m}} \phi_k(m_i)P(k) \propto \prod_{m_i \in \vec{m}} \phi_k(m_i) \tag{5}$$

where the new prescription is denoted by a set of medicines $\vec{m}$, $P(k|\vec{m})$ is the prevalence of treatment method k given the prescription, $P(m_i|k)$ is prevalence of medicine m_i given treatment method k which is equal to $\phi_k(m_i)$ in previous subsection and $P(k)$ is the prior of treatment method k, which can be treated as a constant.

To predict the function of a prescription, a naive approach is using the medicines vector to represent a prescription:

$$\vec{m} = \{m_1, \ldots, m_i, \ldots, m_n\} \tag{6}$$

where each m_i is a binary indicator, if a prescription contains m_i, it is 1, otherwise 0.

We can use the posterior vector as the feature vector of a prescription:

$$\theta_{\vec{m}} = \{P(1|\vec{m}), \ldots, P(i|\vec{m}), \ldots, P(K|\vec{m})\} \tag{7}$$

where each $P(i|\vec{m})$ is the prevalence of treatment methods in Eq. (5).

We can also use the combination of $\vec{m}$ and $\theta_{\vec{m}}$ as features to predict a prescription's function like what [23,24] do when building short text classifiers. The resulting feature space is obtained by appending $\theta_{\vec{m}}$ to $\vec{m}$ as follows:

$$\vec{m} \cup \theta_{\vec{m}} = \{m_1, \ldots, m_i, \ldots, m_n, P(1|\vec{m}), \ldots, P(1|\vec{m}), \ldots, P(K|\vec{m})\} \tag{8}$$

The prediction of a prescription's function (treatment methods) is a multi-label classification problem. We use a set of multiple one-vs-rest classifiers to train and test our model, the classifiers are SVM, Logistic Regression, Adaboost, C4.5, Bayes Network and Random Forest, which are commonly used in data mining.

4. Results

We collected 3090 real world clinical cases made by famous elder TCM masters from literatures in Chinese Knowledge Center for Engineering Science and Technology (CKCEST).[3] As an example, a clinical case is shown in Fig. 3. It describes the patient's chief complaint and history, how a doctor diagnoses and prescribes and the doctor's remark on the case. Texts marked red are identified as symptoms in TCM ontology, and texts marked blue are herbal medicines in *TCM MeSH* which form a prescription.

4.1. Setup

In syndrome labeling step, we set the threshold value of mapping symptoms to syndromes to be 2, i.e., syndromes which have been mapped less than two times will not be selected as candidate syndromes.

In topic discovery step, we set the number of topics to be the number of treatment methods in the clinical data set. There are 37 syndromes like "interior heat syndrome", "wind syndrome" and "syndrome of blood stasis" in TCM ontology, 29 among them have symptoms in their categories, we map 29 syndromes to 29 corresponding treatment methods like "heat-clearing", "wind-relieving" and "blood-regulating". After syndrome labeling, there are $K = 23$ treatment methods left in the data set. And the size of

[3] http://zcy.ckcest.cn/MedicalRecord/index.

清热 (heat-clearing) 理血 (blood-regulating) 治风 (wind-relieving)

大黄(Rhubarb) 枳实(Immature Orange Fruit) 桃仁(Peach Seed) 甘草(Liquorice Root) 半夏(Pinellia Tuber) 茯苓(Poria) 远志(Milkwort Root) 钩藤(Gambir plant Nod) 竹沥(Liquor Phyllostachys) 石菖蒲(Grassleaf Sweetflag Rhizome) 陈皮(Dried Tangerine Peel) 芒硝(Sodium Sulfate)

Fig. 4. The prescription with treatment method labels in Fig. 3. The first row are treatment method labels, the second row are medicines.

Table 1
Topics learned from the clinical cases.

Wind-relieving	Probability	Phlegm-expelling	Probability
Gambir plant Nod	0.0355	Pinellia Tuber	0.0509
Tall Gastrodia Tuber	0.0257	Poria	0.0302
Stiff Silkworm	0.0235	Dried Tangerine Peel	0.0271
Scorpion	0.0229	Gambir plant Nod	0.0260
Rehmannia Glutinosa	0.0226	Bile Arisaema	0.0219
Debark Peony Root	0.0201	Turmeric Root Tuber	0.0198
Unprocessed Rehmannia Root	0.0182	Rhizoma Pinelliae Praeparatum	0.0198
Chinese Angelica	0.0179	Baical Skullcap Root	0.0167
Liquorice Root	0.0179	Citrus Reticulata	0.0167
Divaricate Saposhnikovia Root	0.0176	Glehnia Littoralis	0.0167
dampness-dispelling	Probability	Menstruation and childbirth	Probability
Poria	0.0437	Chinese Angelica	0.0514
White Atractylodes Rhizome	0.0398	Sichuan Lovage Rhizome	0.0346
Alisma Orientale	0.0268	Red Peony Root	0.0330
Dried Tangerine Peel	0.0240	Salvia Root	0.0265
Liquorice Root	0.0240	Cyperus Rotundus	0.0226
Radix Codonopsis	0.0195	Cortex Moutan	0.0220
Plantain Seed	0.0190	Chinese Thorowax Root	0.0210
Pinellia Tuber	0.0176	Peach Seed	0.0207
Ginger	0.0156	Poria	0.0207
Atractylodes Rhizome	0.0148	Debark Peony Root	0.0184
Blood-regulating	Probability	Exterior-releasing	Probability
Chinese Angelica	0.0377	Liquorice Root	0.0380
Salvia Root	0.0367	Bitter Apricot Seed	0.0290
Milkvetch Root	0.0288	Ginger	0.0243
Safflower	0.0280	Poria	0.0216
Sichuan Lovage Rhizome	0.0275	Pinellia Tuber	0.0214
Peach Seed	0.0263	Ephedra	0.0205
Red Peony Root	0.0246	Cassia Twig	0.0204
Liquorice Root	0.0212	Dried Tangerine Peel	0.0144
Medicinal Cyathula Root	0.0181	Debark Peony Root	0.0143
Debark Peony Root	0.0160	Radix Glycyrrhizae Preparata	0.0134

medicines' set is 883. We set the hyper-parameters $\alpha = 50/K$ for each topic in each prescription before incorporating label information and $\beta = 0.1$ as suggested in [22]. The number of iterations is set to be 1000.

In function prediction step, we use the prescriptions with treatment method labels in the labeling step as training set and test set, then use a set of multiple one-vs-rest classifiers (SVM, Logistic Regression, Adaboost, C4.5, Bayes Network and Random Forest) and compare classification performance of three cases, the first is medicine features only (883 features), the second is topic features only (23 features) and the third is combining medicine features and topic features (906 features). The prediction is a 23-class multi-class, multi-label classification problem. We use precision P, recall R and F-measure F as performance measures. The measures are defined as follows:

$$P = \frac{N_c}{N_p}, \quad R = \frac{N_c}{N_r}, \quad F = \frac{2 \times P \times R}{P + R}$$

where N_c is the total number of correct treatment method labels predicted by classifiers in the test set, N_p is the total number of labels predicted by classifiers and N_r is the total number of real labels. We perform 5-fold cross validation (CV), i.e., train classifiers on 2472 clinical cases and test on the remaining 618 cases in 5 runs, and estimate the average performance. We use LibSVM [25] as SVM implementation and Logistic, AdaBoostM1, J48 (C4.5), BayesNet, RandomForest Class in Weka [26] as others' implementation. We tuned SVM's shared cost parameter $C = 100$.

4.2. Experimental result

Syndrome Labeling. We now show the syndrome labeling and treatment method determining result. Taking the clinical case in Fig. 3 as an example, the symptoms marked red are mapped to syndromes "interior heat syndrome", "wind syndrome" and "syndrome of blood stasis" with map count 2, 5 and 5 respectively, which means 2 of them are in "interior heat syndrome" category, 5 of them are in "wind syndrome" category and 5 of them are in 'syndrome of blood stasis" category. They are selected as syndrome labels with map probability larger than the threshold. After labeling syndromes, the treatment methods "heat-clearing", "wind-relieving" and "blood-regulating" are determined by syndrome-treatment method connection in TCM ontology, as shown in Fig. 4, which is quite close to the real situation in this case.[4]

Topic discovering. The 23 topics learned from the clinical cases are shown in Tables 1–4. We show top ten herbal medicines with posterior probability in each topic. Most of them (92.17%) can be

[4] All prescriptions with labels are available at https://github.com/yao8839836/formulae/tree/master/formulae/src/file/prescriptions.

Table 2
Topics learned from the clinical cases.

Tranquillizing	Probability	Heat-clearing	Probability
Milkwort Root	0.0311	Unprocessed Rehmannia Root	0.0277
Liquorice Root	0.0294	Liquorice Root	0.0268
Debark Peony Root	0.0271	Baical Skullcap Root	0.0240
Poria	0.0244	Cortex Moutan	0.0193
Salvia Root	0.0244	Debark Peony Root	0.0186
Acorus calamus	0.0201	Common Anemarrhena Rhizome	0.0181
Indianbread with Pine	0.0187	Rehmannia Glutinosa	0.0165
Rehmannia Glutinosa	0.0181	Golden Thread	0.0165
Unprocessed Rehmannia Root	0.0181	Dwarf Lilyturf Tuber	0.0139
Radix Glycyrrhizae Preparata	0.0164	Weeping Forsythia Capsule	0.0138
Tonifying	Probability	Digestant and Masses Disintegrating	Probability
White Atractylodes Rhizome	0.0305	Dried Tangerine Peel	0.0317
Poria	0.0289	White Atractylodes Rhizome	0.0304
Milkvetch Root	0.0274	Liquorice Root	0.0265
Liquorice Root	0.0247	Orange Fruit	0.0252
Radix Codonopsis	0.0245	Debark Peony Root	0.0252
Debark Peony Root	0.0231	Common Aucklandia Root	0.0212
Chinese Angelica	0.0223	Chinese Thorowax Root	0.0212
Common Yam Rhizome	0.0201	Poria	0.0199
Prepared Rehmannia Root	0.0197	Baical Skullcap Root	0.0186
Rehmannia Glutinosa	0.0185	Ginger	0.0186
Treating abscess and ulcer	Probability	Astringent	Probability
Honeysuckle Flower	0.0314	Common Yam Rhizome	0.0387
Chinese Angelica	0.0279	Milkvetch Root	0.0329
Liquorice Root	0.0243	*Dodder Seed*	0.0271
Milkvetch Root	0.0226	Chinese Magnoliavine Fruit	0.0271
Weeping Forsythia Capsule	0.0221	Oyster Shell	0.0232
Red Peony Root	0.0208	Liquorice Root	0.0213
Dandelion	0.0204	Spine Date Seed	0.0213
Cortex Moutan	0.0195	*Prepared Rehmannia Root*	0.0203
Unprocessed Rehmannia Root	0.0177	Radix Codonopsis	0.0184
Rehmannia Glutinosa	0.0173	Radix Glycyrrhizae Preparata	0.0184

Table 3
Topics learned from the clinical cases.

Anthelmintic	Probability	Resuscitation	Probability
Smoked Plum	0.0437	Liquorice Root	0.0278
Poria	0.0375	Poria	0.0263
Areca Seed	0.0313	Turmeric Root Tuber	0.0263
Common Aucklandia Root	0.0282	*Weeping Forsythia Capsule*	0.0253
Rangooncreeper Fruit	0.0282	Bamboo Shavings	0.0227
Ginger	0.0282	*Immature Orange Fruit*	0.0227
Pricklyash Peel	0.0251	*Rhubarb*	0.0206
Debark Peony Root	0.0220	Baical Skullcap Root	0.0196
Rhubarb	0.0189	Pinellia Tuber	0.0186
Liquorice Root	0.0189	Cow-Bezoar	0.0175
Purgation	Probability	Dryness-relieving	Probability
Rhubarb	0.0218	Dwarf Lilyturf Tuber	0.0519
Radix Et Rhizoma rhei	0.0188	Figwort Root	0.0400
Officinal Magnolia Bark	0.0188	Unprocessed Rehmannia Root	0.0321
Immature Orange Fruit	0.0188	*Dendrobium*	0.0281
Common Aucklandia Root	0.0157	Jade	0.0242
Lotus Leaf	0.0157	Lily Bulb	0.0202
Sodium Sulfate	0.0157	Fragrant Solomonseal Rhizome	0.0202
Dried Tangerine Peel	0.0157	Peppermint	0.0202
Snakegourd Root	0.0157	Loquat Leaf	0.0163
Chinese Dwarf Cherry Seed	0.0157	Cortex Moutan	0.0163
Harmonizing	Probability	Qi-regulating	Probability
Debark Peony Root	0.0500	Debark Peony Root	0.0340
Chinese Thorowax Root	0.0423	Liquorice Root	0.0331
Liquorice Root	0.0375	Cassia Twig	0.0331
Dried Tangerine Peel	0.0346	Poria	0.0321
Ginger	0.0279	Radix Glycyrrhizae Preparata	0.0312
White Atractylodes Rhizome	0.0241	Ginger	0.0302
Pinellia Tuber	0.0222	Pinellia Tuber	0.0265
Common Aucklandia Root	0.0222	Dried Tangerine Peel	0.0237
Orange Fruit	0.0212	Orange Fruit	0.0227
Golden Thread	0.0193	Common Aucklandia Root	0.0208

Table 4
Topics learned from the clinical cases.

Warming interior	Probability	Summerheat-dispelling	Probability
Ginger	0.0547	Fortune Eupatorium Herb	0.0269
Liquorice Root	0.0430	Shrub Chastertree Fruit	0.0269
Radix Codonopsis	0.0361	Chinese Mosla	0.0182
White Atractylodes Rhizome	0.0361	Lotus Leaf	0.0182
Radix Glycyrrhizae Preparata	0.0323	Coix Seed	0.0095
Fresh Ginger	0.0323	*Malaytea Scurfpea Fruit*	0.0095
Pinellia Tuber	0.0303	*Dodder Seed*	0.0095
Milkvetch Root	0.0284	Alisma Orientale	0.0095
Cassia Twig	0.0274	Green Tangerine peel	0.0095
Dried Ginger	0.0254	*Mantis Egg-Case*	0.0095
Exterior-interior dual releasing	**Probability**	**Treating Ear Nose Throat**	**Probability**
Chinese Thorowax Root	0.0892	Xanthium Sibiricum	0.0465
Pinellia Tuber	0.0643	Siberian Cocklebur Fruit	0.0388
Baical Skullcap Root	0.0501	Biond Magnolia Flower	0.0312
Liquorice Root	0.0466	Heartleaf Houttuynia Herb	0.0160
Chinese Date	0.0466	*Asiatic Cornelian Cherry Fruit*	0.0160
Ginger	0.0323	*Amber*	0.0160
Fresh Ginger	0.0288	*White Mulberry Root-Bark*	0.0160
Radix Codonopsis	0.0181	Centipede	0.0160
Amur Cork Tree	0.0181	Sal Ammoniac	0.0084
Sweet Wormwood Herb	0.0146	Barbated Skullcup Herb	0.0084
Depressed Liver Relieving	**Probability**		
Turmeric Root Tuber	0.0420		
Chinese Thorowax Root	0.0173		
Cyperus Rotundus	0.0173		
Cape Jasmine Fruit	0.0173		
Liquorice Root	0.0173		
Common Aucklandia Root	0.0173		
White Atractylodes Rhizome	0.0091		
Rhizoma Pinelliae Praeparatum	0.0091		
Poria	0.0091		
Sichuan Lovage Rhizome	0.0091		

validated in *TCM MeSH* and TCM textbooks, italicized medicines are not in prescriptions of corresponding function category in *TCM MeSH* and textbooks. We can see most of the medicines are validated, others (italicized) could be useful complements to the categories or used together with medicines in the corresponding function category.

Function Prediction. Table 5 shows the classification performance. From the table, we can see that SVM and Bayes Network produce the highest F-Measure scores among 6 classifiers, and Bayes Network achieves the highest F-Measure score 0.5113 on the combined feature space. When using medicine vector features, the result is not good, only SVM and Bayes Network perform relatively well. When utilizing topics features, high precision can be achieved, but the recall is not satisfactory, which means the posterior probability can highlight the most possible treatment methods labels, but ignores other labels. When combining the two feature spaces, the predictive abilities of functions (treatment methods) are improved over medicine features. For F-Measure, the improvement is statistically significant by a 2-tailed paired *t*-test at 95% confidence under all classifiers. The treatment patterns are validated, and can be exploited to understand the TCM clinical data better.

4.3. Discussion

From experimental results, we can see that our method can automatically label TCM clinical cases by syndrome labels, which is useful for clinical case classification and organization; our method can discover medicine usage patterns from a large number of clinical records for each syndrome, which is helpful for

Table 5
Average classification performance of 6 classifiers on three feature spaces.

Classifier	SVM			Logistic regression		
Measures	Precision	Recall	F-measure	Precision	Recall	F-measure
Medcines	0.6306	0.4000	0.4893	0.3021	0.4613	0.3651
Topics	0.7944	0.2343	0.3618	0.7224	0.2965	0.4204
Medicines + topics	0.6470	0.4129	0.5040	0.3091	0.4673	0.3720
	Adaboost			C4.5		
Medcines	0.6675	0.2186	0.3288	0.5448	0.3838	0.4503
Topics	0.6878	0.3379	0.4531	0.6628	0.3532	0.4607
Medicines + topics	0.6944	0.3496	0.4646	0.5609	0.4246	0.4832
	BayesNet			RandomForest		
Medcines	0.5871	0.4367	0.5007	0.6471	0.2775	0.3884
Topics	0.4810	0.4325	0.4553	0.6755	0.3418	0.4539
Medicines + topics	0.5119	0.5109	0.5113	0.6765	0.2815	0.3975

summarizing experiences of TCM doctors; our method can improve prescription function prediction by using medicine co-occurrence information rather than using medicines only, which could provide some suggestions for prescribing. However, our method has some limitations and could be improved.

In syndrome labeling step, we simply use text matching to identify symptoms and medicines. This is a naive approach, which could be improved by named entity recognition [27] and entity linking [28].

It is possible to set the number of topics to a number different from the number of unique treatment methods by using more sophisticated partially supervised topic models [29].

Although we can improve the function prediction performance by using topic features, the result is not very satisfactory. The performance could be improved further by considering medicine information in TCM knowledge. For instance, we can use medicine's function class and description in *TCM MeSH* as features. Additionally, we can also utilize the dosage of each medicine in a prescription.

Our framework can be directly applied to analyze large scale TCM clinical cases in hospitals. The labeling process and topic model could be easily parallelized in MapReduce.

5. Conclusion

This paper has presented a novel framework for mining TCM clinical cases, which consists of syndrome labeling and treatment methods determining, topic discovering and prescription function predicting. The proposed framework exploits data-driven statistical methods and human knowledge to mine clinical data, which could extract useful treatment patterns. Results on real world data set show the effectiveness of our framework. The framework could help doctors diagnose and prescribe after observing symptoms and might be useful to illuminate some further clinical research.

In future work, we plan to update the taxonomy-style ontology (*TCM MeSH*) to more advanced knowledge base which contains more useful relationships, to improve the syndrome labeling and treatment methods determining, and improve prescription function prediction as mentioned in discussions.

Conflict of Interest statement

The authors declare that there is no conflict of interests regarding the publication of this article.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61572434), China Knowledge Centre for Engineering Sciences and Technology (No. CKCEST-2015-2-5) and Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP) (20130101110136).

References

[1] G. Nestler, Traditional chinese medicine, Med. Clin. North Am. 86 (1) (2002) 63–73.
[2] Prescription dictionary [in Chinese], 2014. <http://books.google.com.hk/books?id=0hfBngEACAAJ>.
[3] X. Zhou, Y. Peng, B. Liu, Text mining for traditional chinese medical knowledge discovery: a survey, J. Biomed. Inform. 43 (4) (2010) 650–660.
[4] Z. Deng, Formulae of Chinese Medicine, China Press of Traditional Chinese Medicine, 2008 [in Chinese].
[5] A. Lu, C. Lu, M. Jiang, T. Wei, M. Song, G. Zheng, J. Zhan, Exploring Li-Fa-Fang-Yao rules of major depressive disorder in traditional chinese medicine through text mining, in: 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE, 2013, pp. 460–464.
[6] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, L. Hua, Data mining in healthcare and biomedicine: a survey of the literature, J. Med. Syst. 36 (4) (2012) 2431–2448.
[7] Y. Feng, Z. Wu, X. Zhou, Z. Zhou, W. Fan, Knowledge discovery in traditional chinese medicine: state of the art and perspectives, Artif. Intell. Med. 38 (3) (2006) 219–236.
[8] S. Lukman, Y. He, S.-C. Hui, Computational methods for traditional chinese medicine: a survey, Comput. Meth. Prog. Biomed. 88 (3) (2007) 283–294.
[9] X. Zhou, S. Chen, B. Liu, R. Zhang, Y. Wang, P. Li, Y. Guo, H. Zhang, Z. Gao, X. Yan, Development of traditional chinese medicine clinical data warehouse for medical knowledge discovery and decision support, Artif. Intell. Med. 48 (2) (2010) 139–152.
[10] N.L. Zhang, S. Yuan, T. Chen, Y. Wang, Latent tree models and diagnosis in traditional chinese medicine, Artif. Intell. Med. 42 (3) (2008) 229–245.
[11] G. Zheng, M. Jiang, C. Lu, A. Lu, Prescription analysis and mining, in: Data Analytics for Traditional Chinese Medicine Research, Springer, 2014, pp. 97–109.
[12] J.-j. Lu, J.-k. Pan, H.-m. Li, J.-y. Yang, B.-q. Pan, J. Liu, Research on the component law of chinese medicine for gout and the development of new recipes through unsupervised data mining methods, in: 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE, 2013, pp. 198–201.
[13] H. Yang, J. Chen, S. Tang, Z. Li, Y. Zhen, L. Huang, J. Yi, New drug r&d of traditional chinese medicine: role of data mining approaches, J. Biol. Syst. 17 (03) (2009) 329–347.
[14] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[15] Z. Huang, X. Lu, H. Duan, Latent treatment pattern discovery for clinical processes, J. Med. Syst. 37 (2) (2013) 1–10.
[16] A. Van Esbroeck, C.-C. Chia, Z. Syed, Heart rate topic models, in: Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.
[17] X.-p. Zhang, X.-z. Zhou, H.-k. Huang, Q. Feng, S.-b. Chen, B.-y. Liu, Topic model for chinese medicine diagnosis and prescription regularities analysis: case on diabetes, Chin. J. Integr. Med. 17 (2011) 307–313.
[18] Z. Jiang, X. Zhou, X. Zhang, S. Chen, Using link topic model to analyze traditional chinese medicine clinical symptom-herb regularities, in: 2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom), IEEE, 2012, pp. 15–18.
[19] V. Chenthamarakshan, P. Melville, V. Sindhwani, R.D. Lawrence, Concept labeling: building text classifiers with minimal supervision, in: IJCAI Proceedings-International Joint Conference on Artificial Intelligence, vol. 22, 2011, p. 1225.
[20] L. Wu, Chinese Traditional Medicine and Materia Medical Subject Headings, Chinese Medical Ancient Books Publishing, Beijing, 1996.
[21] D. Ramage, D. Hall, R. Nallapati, C.D. Manning, Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 – Volume 1, Association for Computational Linguistics, 2009, pp. 248–256.
[22] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Nat. Acad. Sci. 101 (suppl. 1) (2004) 5228–5235.
[23] X.-H. Phan, L.-M. Nguyen, S. Horiguchi, Learning to classify short and sparse text & web with hidden topics from large-scale data collections, in: Proceedings of the 17th International Conference on World Wide Web, ACM, 2008, pp. 91–100.
[24] M. Chen, X. Jin, D. Shen, Short text classification improved by learning multi-granularity topics, in: IJCAI, Citeseer, 2011, pp. 1776–1781.
[25] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (3) (2011) 27.
[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The weka data mining software: an update, ACM SIGKDD Explor. Newslett. 11 (1) (2009) 10–18.
[27] E.F. Tjong Kim Sang, F. De Meulder, Introduction to the conll-2003 shared task: language-independent named entity recognition, Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, vol. 4, Association for Computational Linguistics, 2003, pp. 142–147.
[28] D. Rao, P. McNamee, M. Dredze, Entity linking: finding extracted entities in a knowledge base, in: Multi-source, Multilingual Information Extraction and Summarization, Springer, 2013, pp. 93–115.
[29] D. Ramage, C.D. Manning, S. Dumais, Partially labeled topic models for interpretable text mining, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2011, pp. 457–465.