Journal of Biomedical Informatics 45 (2012) 210–223

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

A framework and its empirical study of automatic diagnosis of traditional Chinese medicine utilizing raw free-text clinical records

Yaqiang Wang [a], Zhonghua Yu [a,*], Yongguang Jiang [b], Yongchao Liu [c], Li Chen [a], Yiguang Liu [a]

[a] Department of Computer Science, Sichuan University, Chengdu, Sichuan 610064, PR China
[b] Department of Preclinical Medicine, Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan 610075, PR China
[c] Medical College, Beihua University, Jilin, Jilin 132013, PR China

ARTICLE INFO

Article history:
Received 17 June 2011
Accepted 25 October 2011
Available online 10 November 2011

Keywords:
Automatic diagnosis
Traditional Chinese medicine
Raw free-text clinical records
Natural language processing
Text mining

ABSTRACT

Automatic diagnosis is one of the most important parts in the expert system of traditional Chinese medicine (TCM), and in recent years, it has been studied widely. Most of the previous researches are based on well-structured datasets which are manually collected, structured and normalized by TCM experts. However, the obtained results of the former work could not be directly and effectively applied to clinical practice, because the raw free-text clinical records differ a lot from the well-structured datasets. They are unstructured and are denoted by TCM doctors without the support of authoritative editorial board in their routine diagnostic work. Therefore, in this paper, a novel framework of automatic diagnosis of TCM utilizing raw free-text clinical records for clinical practice is proposed and investigated for the first time. A series of appropriate methods are attempted to tackle several challenges in the framework, and the Naïve Bayes classifier and the Support Vector Machine classifier are employed for TCM automatic diagnosis. The framework is analyzed carefully. Its feasibility is validated through evaluating the performance of each module of the framework and its effectiveness is demonstrated based on the precision, recall and F-Measure of automatic diagnosis results.

1. Introduction

Traditional Chinese medicine (TCM) is becoming a complementary medical theory to western medicine, and it has been gradually accepted and used all around the world [1–3]. Moreover, Data Mining and Machine Learning methods have been applied more than ever before in the field of TCM in recent years to capture regularities from the experience accumulated in the past thousands of years [4–6] to support the TCM experts in clinical research and decision-making.

Most existing work focused on establishing a TCM diagnosis expert system based on manually collected, structured and/or normalized datasets (known as well-structured datasets or experimental datasets) [7–9]. However, the existing work of automatic diagnosis systems, or rather the established TCM expert systems, utilizing the well-structured datasets could not be directly and effectively applied to clinical practice, due to the big difference between the well-structured datasets and the raw free-text clinical records (FCRs). As the basis and the most important reference resources for clinical diagnosis, raw TCM FCRs are unstructured and they are denoted by TCM doctors without the support of authoritative editorial board in their routine diagnostic work. The same concept in raw TCM FCRs may be described in several terms (e.g. the phenomena of "one symptom with different names" [10]) owing to different experience and background of TCM doctors.

Structuring and normalizing these raw TCM FCRs manually are tedious, time consuming, and costly, and at the same time, it would also result in error-prone. Therefore, a huge volume of TCM FCRs, which contains a larger amount of information than well-structured datasets, could not be efficiently and effectively utilized [6]. Consequently, a new framework of automatic diagnosis of TCM utilizing raw FCRs directly demands to be developed for clinical practice. The framework would not only offer a referable way to automatically process the raw TCM FCRs for TCM researchers, but also automatically and effectively guide the TCM practitioners in clinical diagnosis processes.

Four main challenges in such a framework have to be tackled in advance:

(1) How to automatically structure the raw TCM FCRs, e.g. recognizing symptom names from the raw TCM FCRs.
(2) How to process and use the other information contained in raw TCM FCRs, i.e. the fragments except the symptom names

* Corresponding author.
E-mail addresses: wangyaq2204_cn@hotmail.com (Y. Wang), yuzhonghua@scu.edu.cn (Z. Yu), cdtcm@163.com (Y. Jiang), chinalord@hotmail.com (Y. Liu), cl@scu.edu.cn (L. Chen), lygpapers@yahoo.com.cn (Y. Liu).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.10.010

in raw TCM FCRs which may contain the information about attack time of the symptoms, western medical metrics, environment conditions of the symptoms caused, etc. and could support the clinical diagnosis.
(3) How to obtain high quality and valuable clinical evidence for clinical diagnosis, i.e. normalizing recognized symptom names, processing the fragments except the symptom names in the transcripts into segments which could contain semantic unit, and selecting features (i.e. the normalized symptom names and the other information).
(4) How to use the selected (or filtered) features to automatic diagnosis.

There has been no researcher attempting to solve these problems systematically [6]. Hence, in this paper, focusing our attention on the characteristics of the raw TCM FCRs (introduced in Section 2), a novel framework of automatic diagnosis of TCM utilizing raw FCRs for clinical practice is proposed (in Section 3) and investigated (in Section 4). Counting on Natural Language Processing, Data Mining and Machine Learning methods, a series of methods are tried in order to tackle the challenges existing in the framework and investigate the feasibility and effectiveness of the framework.

2. The characteristics of raw TCM free-text clinical records

The transcripts of patients' symptoms in raw TCM FCRs are different a lot from the common texts which are noted by normal Chinese language. The transcripts are denoted by TCM doctors according to the descriptions of physical conditions dictated by patients and the results after the four basic diagnosis procedures (i.e. inspection, olfaction and auscultation, interrogation and palpation) [11], and they have their own noting styles. The contents of the transcripts in raw TCM FCRs are narrative, classical-Chinese-like, and often nonstandard. Getting a clear understand of these characteristics of the transcripts in raw TCM FCRs is very helpful in finding appropriate ways to solve the challenges existing in the framework of automatic diagnosis of TCM utilizing raw FCRs. Therefore, they are summarized and described as follows.

2.1. Narrative form

Most of the transcripts in raw TCM FCRs are written by TCM doctors in narrative form, i.e. several event descriptions of symptoms are represented in-between the sentences of the transcripts. Taking an example, the transcript "昨日肠鸣, 失气多, 心中不适, 早晨大便提早, 头昏, 苔薄, 足转筋, 脉细." (Yesterday, the patient had borborygmus and more farting, and his/her heart was uncomfortable. In this morning, the patient had a bowel movement earlier than before and felt dizziness. The patient is coated tongue thin, feet going into spasms and pulse fine.), in which "昨日肠鸣, 失气多, 心中不适" (Yesterday, the patient had borborygmus and more farting, and his/her heart was uncomfortable.) and "早晨大便提早, 头昏" (In this morning, the patient had a bowel movement earlier than before and felt dizziness.) are two event descriptions of symptoms denoted by the TCM doctor according to the physical conditions dictated by the patient.

2.2. Concise and classical-Chinese-like style

The basic theory of TCM has been founded thousands years. Some traditional and specific habits to describe and record the descriptions of symptoms of patients in clinical diagnosis process are handed down. Thus the transcripts often have the concise and classical-Chinese-like style [12], i.e. the words or phrases used in the transcripts are often short and abbreviate – in other words, some characters in a word or some words in a phrase in the transcripts might be dropped or replaced by brief forms, when their brief forms have been clearly understood by TCM doctors. For example, in the transcript "胃脘胀满, 引两胁胀, 早晨4–5点尤甚" (The patient had a sense of gastric cavity distension, this sense caused his/her two flank distended, and it was especially serious between 4 and 5 o'clock in the morning), "引" (cause) and "尤甚" (especially serious) are, respectively, two brief forms of "引起" (cause) and "尤其严重" (especially serious) which might be used normally.

2.3. Nonstandard description

Because different TCM doctors have their own experience and background and the raw TCM FCRs are denoted by TCM doctors on a day to day basis without any unified standard, one concept, especially symptoms, in the transcripts might be represented by TCM doctors in various terms [10]. For instance, "引胸痛" (lead to stethalgia) can be replaced with "引胸疼痛" (lead to stethalgia) as well as "引胸部痛" (lead to stethalgia) by different TCM doctors.

In addition, various kinds of punctuations are used in the transcripts by TCM doctors also without any restrictions. The same punctuation used in different positions in the transcripts might have types of usages, such as the period "." which could be used as a mark of the end of a sentence, a short form of a word or a decimal point.

3. The framework of automatic diagnosis utilizing raw TCM FCRs for clinical practice

The workflow of automatic diagnosis of TCM using well-structured datasets usually consists of two main processes [5]: (1) selecting features (or variables), which is used to discover more valuable clinical evidence from the features contained in the well-structured datasets; (2) training or constructing a classifier (such as Naïve Bayes classifier, Bayesian Network classifier [7], Support Vector Machine [13], etc.) based on the selected features, and then classifying a new inputted well-structured data into syndromes by the trained classifier.

Referring to the workflow of automatic diagnosis of TCM using well-structured datasets and considering the characteristics of raw TCM FCRs, a novel framework of automatic diagnosis of TCM utilizing raw FCRs for clinical practice is designed. It not only includes the two processes mentioned above, but also, more importantly, takes TCM symptom name recognition and normalization processes and other information usage modes into account. Its architecture (see Fig. 1) is primarily composed of four components that:

(1) FCRs analysis module, which is mainly used to recognize the TCM symptom names and handle the problem how other information to be used.
(2) Normalization module, which would provide normalized symptom names for the following modules.
(3) Feature selection module, which is used to evaluate the worth of the extracted features (i.e. recognized symptom names and processed other information) for automatic diagnosis and then filter out the features which have less diagnosis information to improve the performance of automatic diagnosis.
(4) Training and diagnosing module, which involves the training of a diagnosis model based on the filtered features and then using the trained model to predict an appropriate diagnosis result for a new inputted clinical record.

The detailed roles and necessity of these modules in the framework are introduced as follows, and at the same time, several reasonable methods for each module are introduced.

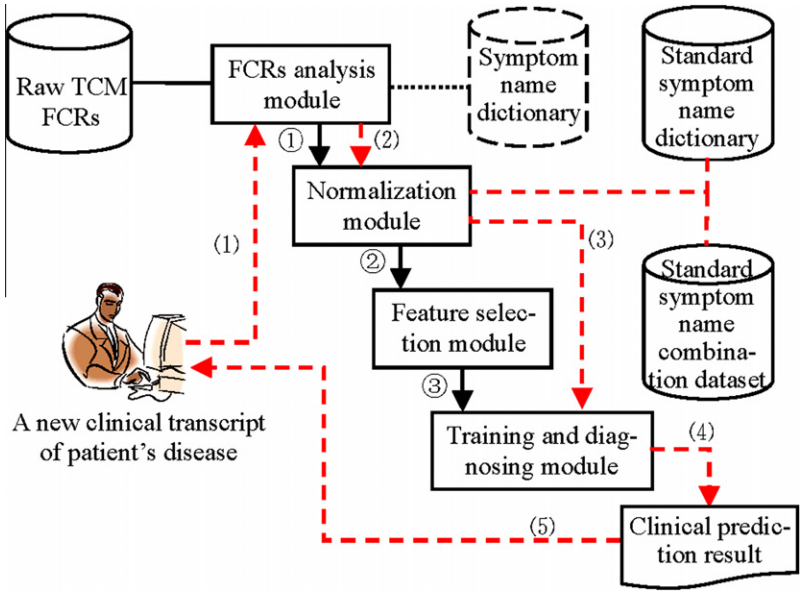

Fig. 1. The architecture of automatic diagnosis utilizing raw TCM FCRs for clinical practice. The workflow ① → ② → ③ is the process of training a model utilizing raw TCM FCRs, and (1) → (2) → (3) → (4) → (5) describes the process of using the trained model to predict a syndrome for a new clinical disease description inputted by a TCM doctor.

3.1. FCRs analysis module

The basic evidence for TCM diagnosing is symptoms of disease, and its descriptions are contained in-between the sentences of the transcripts. And except the symptom names, there is also some other information in the transcripts (i.e. the segments except the symptom names in the transcripts of raw TCM FCRs which may include the information about attack time of the symptoms, western medical metrics, and environment conditions, etc. and could support the clinical diagnosis), and it is used as additional information to further depict the symptoms. Consequently, a FCRs analysis module should be included in the framework. It would perform the tasks that recognizing the TCM symptom names and processing the other information in raw TCM FCRs.

In this paper, the processes of FCRs analysis module are that: (1) segmenting the transcripts in raw TCM FCRs into sub-sentences; (2) recognizing symptom names from the sub-sentences (two types of symptom name recognition method are investigated and introduced below) and processing other information segments based on the characteristics of raw TCM FCRs.

3.1.1. Sub-sentence segmentation

Several methods for sub-sentence segmentation have been proposed in the field of Natural Language Processing, and the problem of sub-sentence segmentation is often treated as recognizing punctuations from sentences (i.e. discovering which punctuations in the sentences are used to mark the end of sub-sentences and which are not). For example, distinguishing what the usage of a period "." in a sentence – is it used to mark the end of a sentence, be the short form of a word, or be a decimal point symbol?

However, there is a slight difference between the ordinary tasks and segmenting the transcripts in that full and half width letter mode punctuations (i.e. Chinese and English punctuations) are used freely in the transcripts of raw TCM FCRs. Therefore, in this paper, they are unified firstly. The Chinese punctuations are changed into their corresponding English mode. Then the GENIA tagger [14–16], which is a widely used tagger for biomedical text, is employed to detect the English punctuations which are used to mark the end of the sub-sentences in the transcripts.

3.1.2. Dictionary-based method

Referring to a symptom name dictionary, the symptom names would be exactly matched from the segmented sub-sentences based on a maximum match algorithm. The procedure of the algorithm is that:

Given a sub-sentence *SubSent* = "$c_1c_2 \ldots c_n$", where c_j is the jth character in *SubSent*, $j \in [1,n]$ and n is the number of characters in *SubSent*, the symptom name(s) in *SubSent* would be recognized through the procedure:

(1) For $m = 1$–n, do:
(2) For $k = 1$–n, do:
(3) If the sub-string "$c_mc_{m+1} \ldots c_k$" of *SubSent* is contained in the symptom name dictionary (in this paper, the symptom name dictionary is manually constructed by TCM experts), then "$c_mc_{m+1} \ldots c_k$" is treated as a potential symptom name. And if "$c_mc_{m+1} \ldots c_{k+1}$" could be matched by a symptom name in the symptom name dictionary, "$c_mc_{m+1} \ldots c_k$" would be abandoned, else "$c_mc_{m+1} \ldots c_k$" would be recognized as a symptom name contained in *SubSent* and m would be set to $k + 1$.
(4) End For.
(5) End For.

Taking an example, through scanning the sub-sentence "腹中有肠鸣音" (a borborygmus in the abdomen), two potential symptom names "肠鸣" (borborygmus) and "肠鸣音" (borborygmus) could be matched according to the symptom name dictionary, but based on the maximum matching algorithm, the longer one "肠鸣音" (borborygmus) (i.e. a maximum matched symptom name in "腹中有肠鸣音") would be regarded as the recognized symptom name in this sub-sentence.

Because the contents of the transcripts are concise, after recognizing the symptom names, most of the segments except the recognized symptom names of the sub-sentences have been individual words, phrases, or semantic fragments. For instance, in a sub-sentence "腹中有肠鸣音" (a borborygmus in the abdomen), after recognizing the symptom name "肠鸣音" (borborygmus) the rest segment "腹中有" (in the abdomen) is already a semantic fragment to indicate an additional diagnosis information. Therefore, the fragments except the recognized symptom names of the sub-

sentences in the transcripts are treated as other information directly.

3.1.3. Bigram-based method

Dictionary-based method could accurately recognized most symptom names which have been included in the symptom name dictionary. However, in clinical practice, a large amount of new symptom names would appear during routine diagnostic work of TCM doctors due to the nonstandard characteristic of symptom names. Thus manually maintaining such a symptom name dictionary is also time-consuming, tedious, and costly.

In addition, some fragments taken as other information in the dictionary-based method could be broken into more detailed additional diagnosis information. In other words, long fragments except the recognized symptom names in the transcripts may cause information loss. For example, the fragment "早晨4–5点尤甚" (especially serious between 4 and 5 o'clock in the morning) could be broken into three kinds of detailed additional diagnosis information that "早晨" (in the morning, which indicate the attack time of the symptom), "4–5点" (between 4 and 5 o'clock, which indicate more accurate attack time of the symptom), and "尤甚" (especially serious, which indicate the incidence degree of the symptom). The best way to cope with this problem is to recognize the detailed information from these fragments or segment these fragments into words. However, due to the classical-Chinese-like characteristic of raw TCM FCRs, automatically segmenting the transcripts into words is dramatically difficult not to mention recognize the detailed information from these fragments. Thus a bigram-based method is investigated, which is a popular method of Natural Language Processing for Chinese [17]. It divides the sub-sentences of a transcript into a list of Chinese character bigrams. Consequently the symptom names or their bigrams are mixed with the bigrams of other information in the bigram list. The hidden symptom names would be recognized as soon as they are normalized in the normalization module based on an investigated method introduced in Section 3.2.2, and the other bigrams except the recognized symptom names would be treated as other information.

3.2. Normalization module

Owing to the nonstandard characteristic of symptom names in raw TCM FCRs, a normalization module is essential to be included in the framework of automatic diagnosis of TCM utilizing raw FCRs. After the processes of this module, the recognized symptom names are normalized, and this refined diagnosis information would improve accuracy of the automatic diagnosis results.

Although symptoms may be described by TCM doctors in several different names due to the different experience and background each TCM doctor has, symptom names which represent the same symptom usually have literal similarity. According to this attribute, several TCM symptom name normalization methods have been proposed in [10], and the literal similarity metric introduced in [10] is more suitable to be used in the framework of automatic diagnosis of TCM utilizing raw FCRs for clinical practice.

Therefore, based on the ideas of literal similarity metric described in [10] and focusing on the two types of outputs generated by dictionary-based and bigram-based method which are introduced in Section 3.1, two types of symptom name normalization methods for clinical practice are designed in this paper. Smith-Waterman distance literal similarity metric (SWD) is used in our experiments. This metric could achieve the best normalization results among literal similarity metrics, and its threshold is set to 0.7 in this paper (because according to the results reported in [10], when the threshold is assigned to 0.7, relatively better precision, recall, and F-Measure would be obtained by SWD). The details of the investigated symptom name normalization methods are described as follows.

3.2.1. Normalizing symptom names generated by dictionary-based method

After the process of dictionary-based FCRs analysis method, each transcript would be converted to a symptom name and other information segment list denoted as $SNOISegList = \{SympSeg_1, \ldots, SympSeg_n, OthInfSeg_1, \ldots, OthInfSeg_m\}$, where $SympSeg_i$ is the ith recognized symptom name in the list, $i \in [0, n]$, n is the number of recognized symptom names in the transcript, and $OthInfSeg_j$ is the jth other information segment detected by dictionary-based method, $j \in [0, m]$ and m is the number of other information segments detected from the transcript. Based on a standard symptom name dictionary ($StdSympDic = \{StdSymp_k$, where $k \in [1, K]$ and K is the number of standard symptom names in $StdSympDic\}$) (in this paper, the standard symptom name dictionary is manually constructed by TCM experts), the recognized symptom names in each $SNOISegList$ would be normalized after the processes introduced below.

(1) For each $SympSeg_i$ in $SNOISegList$ and each $StdSymp_k$ in $StdSympDic$, if $SWD(SympSeg_i, StdSymp_k) \geqslant 0.7$, then $StdSymp_k$ is a potential standard form of $SympSeg_i$, and it would added into a potential normalized symptom name set $PotNormSympNSet_i$ corresponding to $SympSeg_i$.
(2) For each $PotNormSympNSet_i$, if $PotNormSympNSet_i = null$, then $SympSeg_i$ would be added into $PotNormSympNSet_i$ and be regarded as the standard form of itself.
(3) Selecting one normalized symptom name from each generated potential normalized symptom name set, then P potential normalized symptom name combinations could be generated, and the pth potential normalized symptom name combination is denoted as $PotNormSympNComb_p$, where $P = \prod_{i=1}^{n} SizeOfPotNormSympNSet_i$, $p \in [1, P]$ and $SizeOfPotNormSympNSet_i$ is the size of $PotNormSympNSet_i$.
(4) According to the intuition that "if the symptom names have a great possibility to appear simultaneously, they would have a greater chance to appear in the same clinical record in the future", in this paper, the potential normalized symptom name combination which had the greatest possibility to appear simultaneously are chosen as the normalized result. The possibility of each $PotNormSympNComb_p$ is measured by the combination value $CombValue_p$, and

$$CombValue_p = \arg\max_q \frac{|PotNormSympNComb_p \cap StdComb_q|}{|PotNormSympNComb_p \cup StdComb_q|} \quad (3.1)$$

where $StdComb_q$ is the qth standard symptom name combination in a standard symptom name combination dataset (in this paper, the standard symptom name combination dataset is extracted from a standard prescription dataset). If there is a tie, one of the potential normalized symptom name combinations having the highest combination value would be chosen randomly as the normalized result.

3.2.2. Normalizing symptom names hidden in the bigram lists generated by bigram-based method

Each transcript in raw TCM FCRs would be converted to a bigram list by bigram-based method, and the generated bigram lists would include the symptom names or their bigrams which are mixed with the bigrams of other information. In order to recognize the symptom names and distinguish them from the bigrams of other information, a particular and feasible symptom name normalization method is investigated in this paper. The normalization method would automatically recognize the symptom names which are hidden in the bigram lists as soon as normalize these hidden symptom names. The processes of this method are presented below.

Given a transcript $Transcript = \{SubSent_i, i \in [1, n]\}$, where $SubSent_i$ is the ith sub-sentence in *Transcript* and n is the number of sub-sentence in *Transcript*. After the process of bigram-based FCRs analysis method, each sub-sentence in *Transcript* would be segmented into bigrams, denoted as $SubSentBigram_i = \{bigram_{i,j}, j \in [1, m]\}$, where $bigram_{i,j}$ is the jth bigram generated from $SubSent_i$ and m is the number of bigrams in $SubSentBigram_i$.

Then, for $SubSentBigram_i$, neighboring bigrams in it could be combined and merged into one string, thus several possible bigram combination lists would be generated. For example, based on the bigram list "肌肤, 肤甲, 甲错", four possible bigram combination lists would be got ("肌肤, 肤甲, 甲错", "肌肤甲, 甲错", "肌肤, 肤甲错", and "肌肤甲错"). The possible bigram combination lists of $SubSent_i$ are maintained by a set denoted as $BigramCombListSet_i$.

Consequently, through selecting one possible bigram combination list from each combination list set, P candidate splitting forms of *Transcript* would be generated, and the pth candidate splitting form is denoted as $CandSplitForm_p = \{candsplitform_i, i \in [1, n]\}$, where $candsplitform_i$ is a possible bigram combination list chosen from $BigramCombListSet_i$ of $SubSent_i$, $p \in [1, P]$, $P = \prod_{i=1}^{n} SizeOfBigramCombListSet_i$, and $SizeOfBigramCombListSet_i$ is the size of $BigramCombListSet_i$.

For each element seg_t in $CandSplitFormComb_p$ and each $StdSymp_k$ in *StdSympDic* which is a standard symptom name dictionary and mentioned in Section 3.2.1, if $SWD(seg_t, StdSymp_k) \geqslant 0.7$, then seg_t would be treated as a potential symptom name denoted as $PotSympSeg_u$ (because when the similarity is greater than 0.7, the element has a greater possibility to be normalized by the standard symptom name, in other words, it is more like a symptom name), where $t \in [1, T]$, T is the number of elements in $CandSplitFormComb_p$, $u \in [0, U]$ and U is the number of potential symptom names detected in $CandSplitFormComb_p$. And $StdSymp_k$ would be regard as a potential standard form of $PotSympSeg_u$ and added into a potential normalized symptom name set $PotNormSympSegSet_u$ corresponding to $PotSympSeg_u$.

Through selecting one potential normalized symptom names from each potential normalized symptom name sets of $CandSplitFormComb_p$ and based on the P candidate splitting forms, Q potential normalized symptom name combinations could be generated, and the qth potential normalized symptom name combination is denoted as $PotNormSympNComb_q$, where $q \in [1, Q]$, $Q = P\prod_{u=1}^{U} SizeOfPotNormSympNSet_u$, $SizeOfPotNormSympNSet_u$ is the size of $PotNormSympSegSet_u$.

According to the intuition "if the symptom names have a great possibility appearing in the same record, they would have a greater possibility to appear in the same record in the future" which has been mentioned before, the potential normalized symptom name combination which had the greatest possibility to appear simultaneously would be chosen as the normalized result. And the possibility of each $PotNormSympNComb_q$ is measured by the combination value $CombValue_q$, and $CombValue_q$ could be also computed by Eq. (3.1). If there is a tie, one of the potential normalized symptom name combination would be chosen randomly as the normalized result. The corresponding potential symptom names of the best potential normalized symptom name combination would be treated as the recognized symptom names in *Transcript*.

Referring to the representation described in Section 3.1.2, the other elements except recognized symptom names in the corresponding possible bigram combination of the best potential normalized symptom name combination would be re-divided into bigrams. And these bigrams would be treated as other information.

3.3. Feature selection module

After the processes of FCRs analysis module and normalization module, a large amount of features (i.e. normalized symptom names and the other information segments) would be obtained. Feature selection should be follow-onto evaluate the worth of the extracted features and filter out the features which have less diagnosis information to improve the performance of automatic diagnosis. In this paper, three classical feature selection methods (i.e. Document Frequency method, Mutual Information method, and Information Gain method) are investigated.

3.3.1. Document Frequency method

Document Frequency (DF) [18] of a unique feature in raw TCM FCRs is the number of clinical records in which this unique feature occurs. The intuition of this method is that the features could contain enough diagnosis information when their frequencies in raw TCM FCRs reach a predefined threshold, and these features would be more important and informative for diagnosing than the rare features whose frequencies are under the predefined threshold [18].

3.3.2. Mutual Information method

Mutual Information (MI) [18] is a more commonly used method in statistical language modeling for feature selection. Differing from DF, MI is a quantitative measurement of mutual dependence (or correlation) between a feature f and a class label c_i rather than a simple measurement of the frequency of f in raw TCM FCRs (in this paper, c_i is a syndrome label, i.e. a diagnosis result). The feature is more significant if it has stronger correlation with syndrome labels, in other words, the Mutual Information $I(f, c_i)$ between f and c_i is higher. $I(f, c_i)$ is defined as follows.

$$I(f, c_i) = \log \frac{P(f|c_i)}{P(f)P(c_i)} \tag{3.2}$$

where $P(f|c_i)$ is the conditional probability of f given c_i, and $P(f)$ and $P(c_i)$ are, respectively, the prior probabilities of f and c_i in raw TCM FCRs. Then the global worth of f in raw TCM FCRs is evaluated by the following formula.

$$I(f) = \sum_{i=1}^{m} P(c_i) I(f, c_i) \tag{3.3}$$

where m is the number of unique syndrome labels in raw TCM FCRs.

3.3.3. Information Gain method

Differing a lot from DF and MI, Information Gain (IG) [18] is used to measure the quantity of the diagnosis information that a feature contains, and it is often employed as a criterion for evaluating the effectiveness of a feature in the field of Machine Learning [18]. If a feature has stronger predictive role to one particular diagnosis result, then its IG value will be higher, i.e. it contains more diagnosis information and has higher distinguishability. In other words, the TCM doctor could give the diagnosis result as quick and confident as possible when this feature appears. The IG value of a feature f in raw TCM FCRs is defined below.

$$\begin{aligned} IG(f) = & -\sum_{i=1}^{m} P(c_i) \log P(c_i) \\ & + P(f) \sum_{i=1}^{m} P(c_i|f) \log P(c_i|f) \\ & + P(\bar{f}) \sum_{i=1}^{m} P(c_i|\bar{f}) \log P(c_i|\bar{f}) \end{aligned} \tag{3.4}$$

where m is the number of unique syndrome labels in raw TCM FCRs, $P(c_i)$ is the prior probabilities of c_i in raw TCM FCRs, $P(f)$ is the probability of f occurring in raw TCM FCRs and $P(\bar{f})$ is the probability that f does not occur in raw TCM FCRs. $P(c_i|f)$ is the probability of c_i given f and $P(c_i|\bar{f})$ is the probability that c_i occurs without f.

3.4. Training and diagnosing module

As the main part in the framework of automatic diagnosis of TCM utilizing raw FCRs, training and diagnosing module will accomplish the missions that, firstly, it trains a diagnosis model based on the processed raw TCM FCRs, and then when a TCM practitioner inputs a new transcript, the module will predict an appropriate diagnosis result according to the trained model.

Based on manually well-structured datasets, several approaches have been employed in TCM automatic diagnosis research, such as Naïve Bayes classifier and Bayesian Network classifier [7], Support Vector Machine (SVM) [13], etc. In this paper, Naïve Bayesian classifier (NB) and Support Vector Machine classifier (SVM) are re-employed for TCM automatic diagnosis to investigate the availability of the features extracted, normalized and filtered after previous modules and study specialties of the classifiers in automatic diagnosis utilizing raw FCRs. For convenience, in this paper, the transcripts would be converted to a feature vector that each feature in the vector would be directly and uniformly assigned to binary weight (i.e. 1 (if the feature appears in the transcript) or 0 (if the feature does not appear in the transcript)).

3.4.1. Naïve Bayes classifier for TCM diagnosing

Naïve Bayes classifier (NB) is a typical generative model, which emphasizes the importance of prior knowledge and assumes that all features are independent given the class variables, and it has been applied widely in automatic diagnosis [7,19,20]. For TCM automatic diagnosis, NB focuses on finding the diagnosis principles which are contained in raw TCM FCRs, or rather discovering the quantitative relationships between a diagnosis feature vector and one of the syndrome labels.

NB for TCM automatic diagnosis uses the Bayes rule to compute the posterior probability of a syndrome label given a new clinical record, and it will output one syndrome label with highest posterior probability as the diagnosis result, i.e. given a new free-text clinical record *fcr* which could be featured in a vector $F = (f_1, f_2, ..., f_n)$, where n is the number of features in *fcr*. F would be obtained after the processes of FCRs analysis module and normalization module and the features in F would also be filtered based on the results of feature selection module. The posterior probability of each syndrome label s_i given *fcr* can be calculated by

$$P(s_i|fcr) = P(s_i|F) = P(s_i|f_1, f_2, \ldots, f_n) = \frac{P(f_1, f_2, \ldots, f_n|s_i)P(s_i)}{P(f_1, f_2, \ldots, f_n)} \quad (3.5)$$

where $i \in [1, N]$ and N is the number of syndrome labels in raw TCM FCRs for training. Supposing that the features are independent of each other, then

$$P(s_i|fcr) = \frac{P(s_i)\prod_{j=1}^{n} P(f_j|s_i)}{\prod_{j=1}^{n} P(f_j)} \quad (3.6)$$

Because $\prod_{j=1}^{n} P(f_j)$ is same for all s_i, finally the diagnosis result $S_{\max}$ of *fcr* is determined by

2008 SiJunZi Standard TCM Dataset (SJZTCMD):

Prescription table in SJZTCMD:

PID	Prescription Name	Reference	Type of Prescription	Dynasty
19	一阴煎	《景岳全书·新方八阵	汤	明

Presc_Symptom table in SJZTCMD:

PID	Standard Symptom Name
19	潮热
19	脉虚
19	心烦
19	气短

Clinical records dataset (CRD):

Diagnostic Date	Patient Name	Gender	Age	Transcript of patients' diseas	Diagnosis result	Formulas	Drugs
2004/6/3	[redacted]	女	21	小便镜检血尿2-3+, 尿血, 尿频时作, 小腹胀坠, 腰酸, 舌尖暗红, 脉弦.	肝脾气陷	丹栀逍遥散	丹皮, 栀子, 柴胡, 白芍, 茯苓, 薄荷, 旱连草, 白茅根, 生地, 黄芪, 补骨脂, 炒蒲黄, 甘草, 益母草.
2004/6/8	[redacted]	男	38	胃脘胀满, 引两胁胀, 早晨4-5点尤甚, 引胸痛, 容易气郁, 矢气, 心中烧灼感, 不吐酸, 不敢吃冷食, 口苦, 身软, 睡眠差, 小便正常, 苔略厚腻, 脉弦.	寒热错杂	半夏泻心汤	半夏, 党参, 黄芩, 黄连, 炮姜, 广香, 枳壳, 砂仁, 乌贼骨, 台乌, 淮山, 白术, 甘草.
2004/4/19	[redacted]	女	62	昨日肠鸣, 矢气多, 心中不适, 早晨大便提早, 头昏, 苔薄, 足转筋, 脉细.	脾肾阳虚	桂附理中汤	桂枝, 附子, 党参, 白术, 干姜, 广香, 砂仁, 黄芪, 补骨脂, 肉豆蔻, 吴茱萸, 草果, 甘草, 鹿角霜
2008/10/6	[redacted]	男	60	长期从事跑步, 脚跟痛 脚后跟痛, 上肢冷, 耳鸣, 目干, 舌淡, 脉缓 下肢静脉曲张/15年前手术	血虚, 血瘀	桃红四物	桃仁1.5, 红花1, 当归2, 川芎2, 白芍2, 生地2, 黄芪3, 桂枝1.5, 生姜1.5, 大枣1.5, 炙草1 =50/2/60

EVALDATA:

CLINICAL_SYMPTOM_NAME	STANDARD_SYMPTOM_NAME
吐酸	泛酸, 吞酸
尿微黄	小便黄
心中不乐	情绪抑郁
口渴不明显	口渴
腹泻后服西药腹胀	腹泻
双腿皮肤痒	皮肤瘙痒
口黏腻	口腻

Fig. 2. Examples of experimental datasets.

$$S_{\max} = \arg\max_i(P(s_i|fcr)) = \arg\max_i(P(s_i)\prod_{j=1}^{n} P(f_j|s_i)) \quad (3.7)$$

where $P(s_j)$ could be directly estimated from raw TCM FCRs for training, and in order to tackle the issue of data sparsity, i.e. the problem of unseen features which problem the framework has to face up to. Therefore, in this paper, $P(f_j|s_i)$ is estimated based on Lidstone rule which is a smoothing method and shown as follows:

$$P(f_j|s_i) = \frac{N_{f_j,s_i} + \lambda}{N_{s_i} + \lambda B} \quad (3.8)$$

where N_{f_j,s_i} is the number of clinical records containing both f_j and s_i, and N_{s_i} is the number of clinical records whose diagnosis results are si in raw TCM FCRs for training. λ and B are two constant and empirical values. λ is usually small and B is often a very large value (it is often set to the vocabulary or feature size). In our experiment, λ is assigned to 0.5 [21], and the number of features generated after normalizing symptom names hidden in bigram lists in our dataset is approximately 20,000, and the number of features generated by dictionary-based method is comparatively small, thus B is set to 20,000.

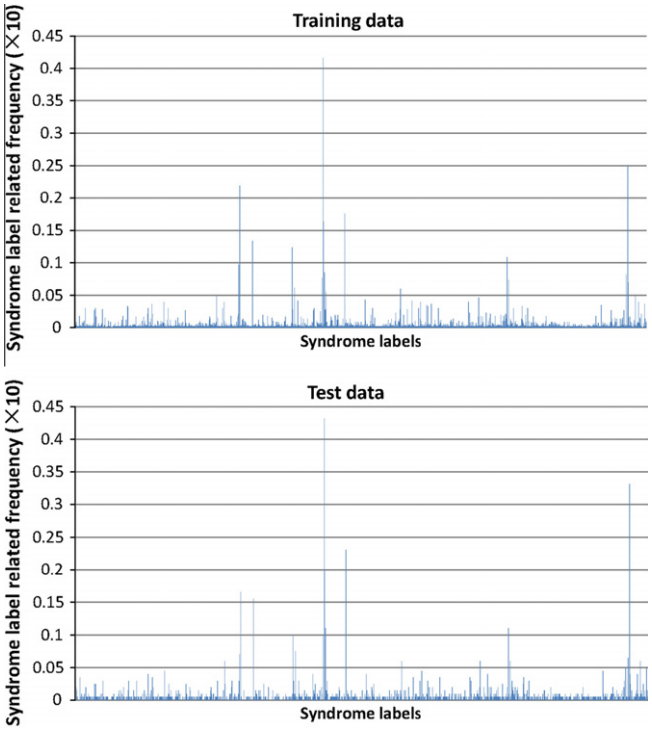

Fig. 3. Related frequency of each syndrome label in training data and test data.

3.4.2. Support Vector Machine classifier for TCM diagnosing

Contrasting with NB, Support Vector Machine (SVM) places emphasis on modeling the posterior $P(s_i|fcr)$ directly, i.e. automatically discovering the global diagnosis regularities from raw TCM FCRs directly.

The task of automatic diagnosis of TCM is a multi-class classification. A straightforward approach is to train a SVM for each syndrome label (i.e. one-against-rest strategy). However, it is not an elegant and excellent approach to solving multi-class classification problem. In our experiment, this problem is made up by LIBSVM [22] which is a well-known SVM tool for multi-class classification. The radial basis function is used as the kernel function of LIBSVM and the other parameters used in LIBSVM are assigned to default values.

3.5. Evaluation metrics

According to the characteristics of raw TCM FCRs (described in Section 2), a novel framework of automatic diagnosis of TCM utilizing raw FCRs has been designed, and several appropriate methods are investigated for each module of the framework. For evaluating the feasibility and effectiveness of the proposed framework, three types of evaluation metrics are designed: (1) for appraising the ability of symptom name recognition, the recognition rate and recognition error rate are exercised; (2) for assessing the performance of normalization module, the normalization precision, recall and F-Measure are tested; (3) the diagnosis precision, recall and F-Measure are defined for comprehensively judging the feasibility and capability of the automatic diagnosis framework of TCM utilizing raw FCRs. These evaluation metrics are described in detail as follows.

3.5.1. Recognition rate and error rate

The recognition rate metric (RR_{rec}) and error rate metric (ERR_{rec}) are used for assessing the ability of symptom name recognition in the framework. Better recognition ability would be obtained, when higher RR_{rec} and lower ERR_{rec} are achieved. RR_{rec} and ERR_{rec} are formulated as follows.

$$RR_{rec} = \frac{|NSRC|}{|NS|} \quad (3.9)$$

$$ERR_{rec} = \frac{|SR| - |NSRC|}{|SR|} \quad (3.10)$$

where $|NSRC|$ is the number of symptom names recognized correctly, $|NS|$ is the number of clinical symptom names in a raw TCM free-text clinical record dataset, and $|SR|$ is the number of symptom names recognized.

In this paper, two ways are trailed for judging the correctness of the recognized symptom names in the process of normalizing symptom names hidden in the bigrams lists. One is that: if a recog-

Table 1
Names, purposes and compositions of the five datasets used in the framework. The related frequencies of syndrome labels in raw TCM FCRs dataset and Inputted test dataset are shown in Fig. 3 respectively.

Name of the dataset	Purpose	Composition
Raw TCM FCRs dataset	Training the classifiers for diagnosing	5979 Records (75% of CRD), chosen from CRD randomly
Symptom name dictionary	Serving as the symptom name dictionary	Symptom names included in EVALDATA, these symptom names could cover all clinical symptom names in CRD
Standard symptom name dataset	Treating as the standard symptom name dictionary	The distinct symptom names appearing in SJZSTCMD
Standard symptom name combination dataset	Regarding as the standard symptom name combination dataset mentioned in the framework	Symptom name combinations which appear in the same record of prescription table of SJZSTCMD
Inputted test dataset	Transcripts in these records are taken as the new clinical transcripts, and their corresponding syndrome labels used to test the automatic diagnosis results	1993 Records (25% of CRD), chosen from CRD randomly

Table 2
Evaluation of the ability of symptom name recognition. The second row is the results under the rough condition that treating the complete or partial matching symptom names as the correct results. The third row is the results under the condition with more strict constraint described in Section 3.5.1.

Method	$\|NSRC\|$	$\|NS\|$	$\|SR\|$	RR_{rec} (%)	ERR_{rec} (%)
DM	39,660	39,660	39,660	100	0
NGBM (roughly)	38,676	39,660	41,756	97.52	7.38
NGBM	27,438	39,660	41,756	69.18	34.29

Table 3
Results of two types of normalization approaches used in normalization module. The bold values are relatively better normalization results achieved by NGDM or NGBM.

Normalization approach	Pre_{norm} (%)	Rec_{norm} (%)	FM_{norm} (%)
NGDM	**73.18**	64.03	**68.30**
NGBM	53.88	**69.18**	60.58

nized symptom name could match the symptom name appearing in that position of current clinical record partially, in other words, it is a sub-string of the symptom name in that position of current clinical record, and then the recognized symptom name is deemed to be correct. Another way is that: a recognized symptom name is correct, if and only if the recognized symptom name perfectly match the symptom name which should appear in that position of current clinical record.

3.5.2. Normalization evaluation metrics

The normalization results are evaluated by normalization precision (Pre_{norm}), recall (Rec_{norm}) and F-Measure (FM_{norm}), and they are defined as follows:

$$Pre_{norm} = \frac{|SNNC|}{|SNN|} \tag{3.11}$$

$$Rec_{norm} = \frac{|SNNC|}{|NSN|} \tag{3.12}$$

$$FM_{norm} = \frac{2 \cdot Pre_{norm} \cdot Rec_{norm}}{Pre_{norm} + Rec_{norm}} \tag{3.13}$$

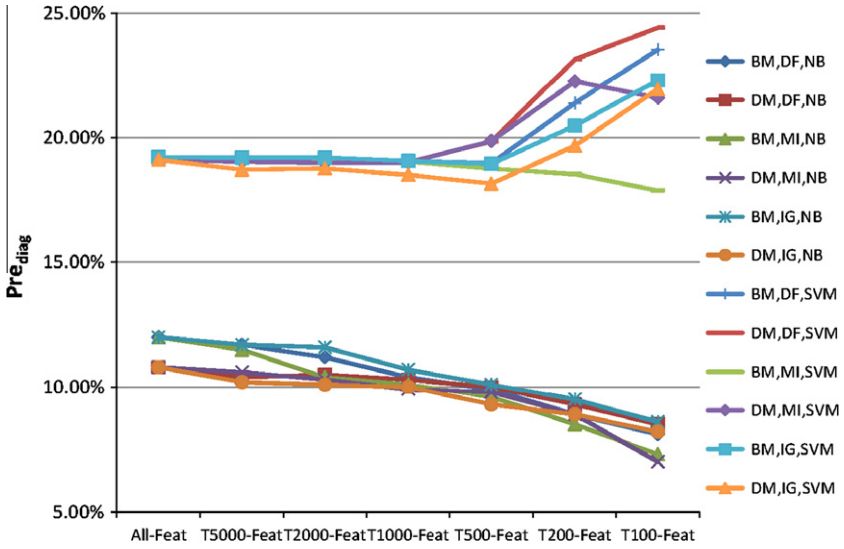

Fig. 4. Results of Pre_{diag} obtained by NB and SVM when different FCRs analyzing methods, normalization methods and feature selection methods are used. TN-Feat means that after feature selection only top N features are used by training and diagnosing module.

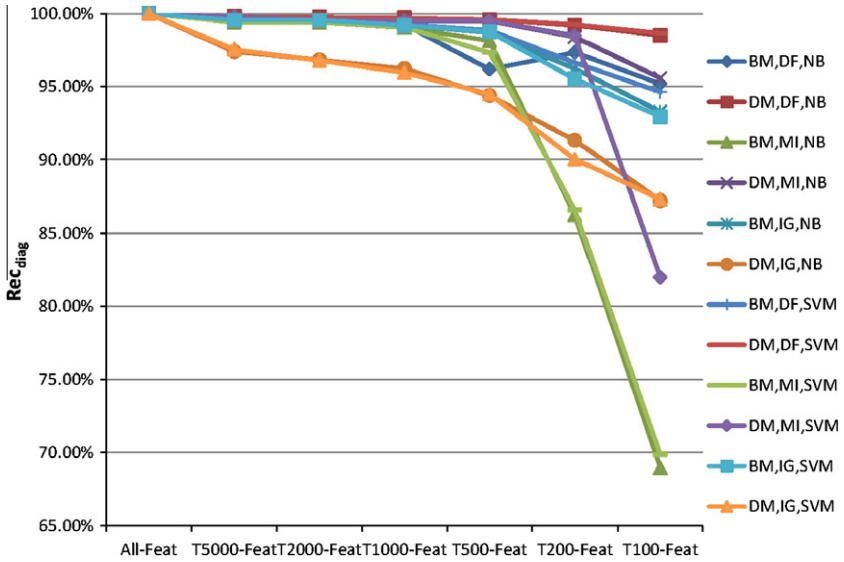

Fig. 5. Results of Rec_{diag} obtained by NB and SVM when different FCRs analyzing methods, normalization methods and feature selection methods are used. TN-Feat means that after feature selection only top N features are used by training and diagnosing module.

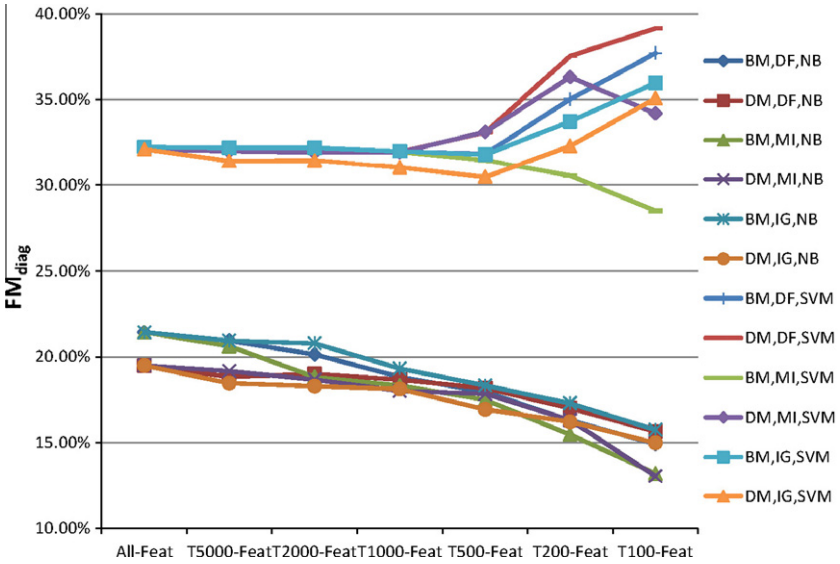

Fig. 6. Results of FM_{diag} obtained by NB and SVM when different FCRs analyzing methods, normalization methods and feature selection methods are used. TN-Feat means that after feature selection only top N features are used by training and diagnosing module.

where $|SNNC|$ is the number of symptom names normalized correctly, $|SNN|$ is the number of symptom names normalized, and $|NSN|$ is the number of nonstandard symptom names should be normalized in a raw TCM free-text clinical record dataset.

3.5.3. Evaluation metrics for automatic diagnosis performance

The diagnosis precision (Pre_{diag}), recall (Rec_{diag}) and F-Measure (FM_{diag}) are used not only for assessing the global performance of the proposed framework but also for evaluating the worth of FCRs analyzing module and normalization module in the framework, and they are defined as follows:

$$Pre_{diag} = \frac{|CDR|}{|DR|} \tag{3.14}$$

$$Rec_{diag} = \frac{|DR|}{|AIR|} \tag{3.15}$$

$$FM_{diag} = \frac{2 \cdot Pre_{diag} \cdot Rec_{diag}}{Pre_{diag} + Rec_{diag}} \tag{3.16}$$

where $|CDR|$ is the number of raw TCM FCRs correctly diagnosed, $|DR|$ is the number of raw TCM FCRs diagnosed, and $|AIR|$ is the number of all inputted raw TCM FCRs.

4. Results

In this section, based on several experimental datasets, the proposed framework is evaluated from several aspects introduced in Section 3.5, and the results are described below.

4.1. Experimental datasets

In this paper, three datasets are used. The first one is the 2008 SiJunZi Standard TCM Dataset (SJZSTCMD). It is a national standard dataset including 4950 prescriptions with 947 distinct symptom names. The second one is a clinical record dataset (CRD) which contains 7972 free-text clinical records with their corresponding diagnosis results, and it is collected by TCM doctors during their routine diagnostic work. There are 4465 distinct clinical symptom names contained in CRD, and they could be normalized to one of the 947 standard symptom names. The last one is a clinical-standard symptom name reference dataset, named EVALDATA. It consists of clinical symptom names with their standard forms and is manually constructed by TCM experts according to CRD and SJZSTCMD. In this paper, it is used to evaluate the results of symptom name normalization. Examples of these datasets are shown in Fig. 2.

In order to simulate the diagnosis procedure of clinical practice to test the performance of the proposed framework, based on SJZSTCMD, CRD, and EVALDATA, five datasets mentioned in Fig. 1 are built, and the detailed introduction are listed in Table 1. For convenience, all numbers in CRD, such as integers, decimals and fractions, are uniformly replaced by the character "*N*" automatically in advance.

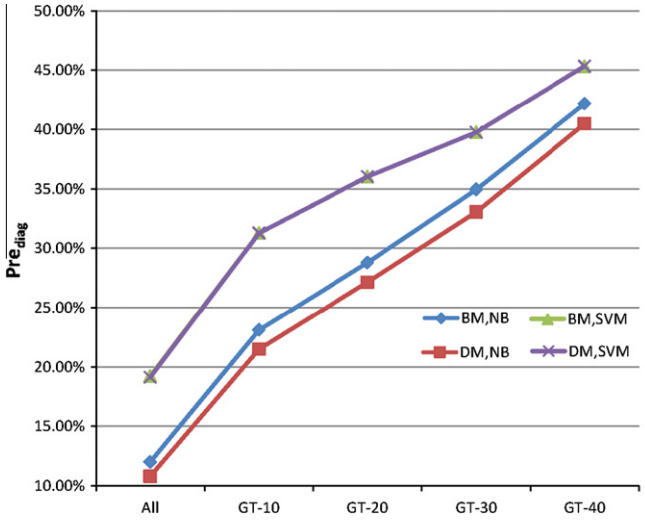

Fig. 7. Comparison of precisions obtained by different classifiers (NB and SVM) when the records which include low frequency syndrome labels (i.e. disease) in CRD are removed and all features are used by training and diagnosing module. In this figure, GT-N means the free-text clinical record in CRD would be used in the experiments when the frequency of its contained syndrome label appearing in raw FCRs is more than N times, and at the same time, the records containing the same syndrome labels are also removed from test data.

4.2. Evaluation of symptom name recognition ability

The abilities of symptom name recognition in the processes of dictionary-based FCRs analysis method (DM) and normalizing symptom names hidden in the bigram lists generated by bigram-based method (NGBM) are evaluated based on the recognition rate metric and error rate metric.

The details of the evaluation results are shown in Table 2. Under the rough measurement, NGBM has a good performance (RR_{rec} = 97.52% and ERR_{rec} = 7.38) on recognizing symptom names from the transcripts. It reveals that extracting the symptom names from raw TCM FCRs without the support of a complementary symptom name dictionary and, at the same time, just using the existing standard dada sources is possible and could be achieved. RR_{rec} of NGBM under the strict condition reduces from 97.52% to 69.18%, and ERR_{rec} of NGBM rises by about 26.91%, the main reason for these results might be caused by the relatively simple ways to determine the potential symptom names and select the best normalization result. It indicates that more robust and advanced methods for recognizing clinical symptom names should be developed and investigated to improve the recognition performance. For example, more linguistic features of the symptom names or the transcripts should be considered to be brought into the recognition methods. And at the same time, these methods should not too much depend on the complementary symptom name dictionary. Thus huge amount of repeated labor to construct a complementary symptom name dictionary for TCM clinical researchers could be cut down.

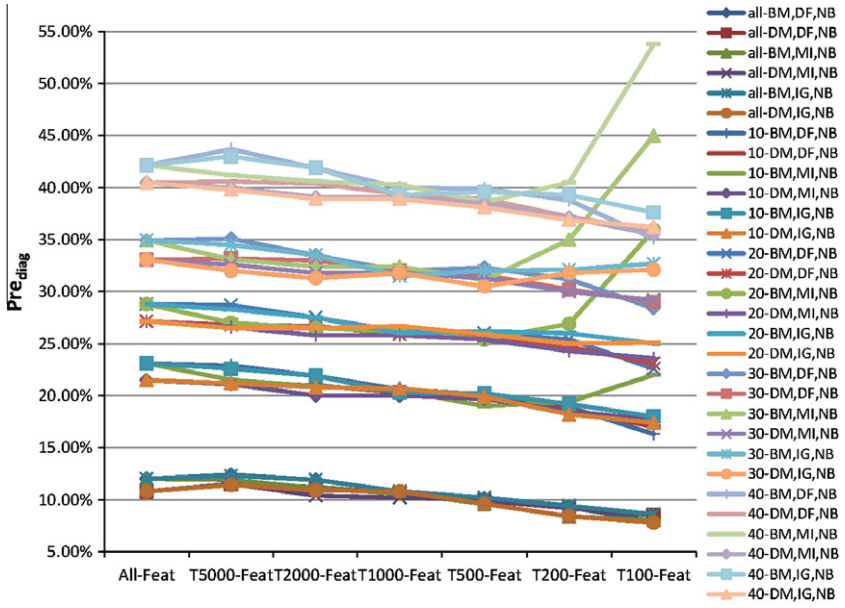

Fig. 8. Results of Pre_{diag} obtained by NB classifier under different situations. Where All-Feat or TN-Feat represents that diagnosing based on all or top N features sorted by the results of different feature selection methods (i.e. DF, MI and IG), and "N-(BM or DM), (DF, MI, or IG), NB" means the experiments perform on the records whose class labels appear more than N times in raw TCM FCRs, use bigram- or dictionary-based method to segment the raw TCM FCRs, Document Frequency threshold, Mutual Information or Information Gain method to select features and utilize Naïve Bayes classifier to diagnose.

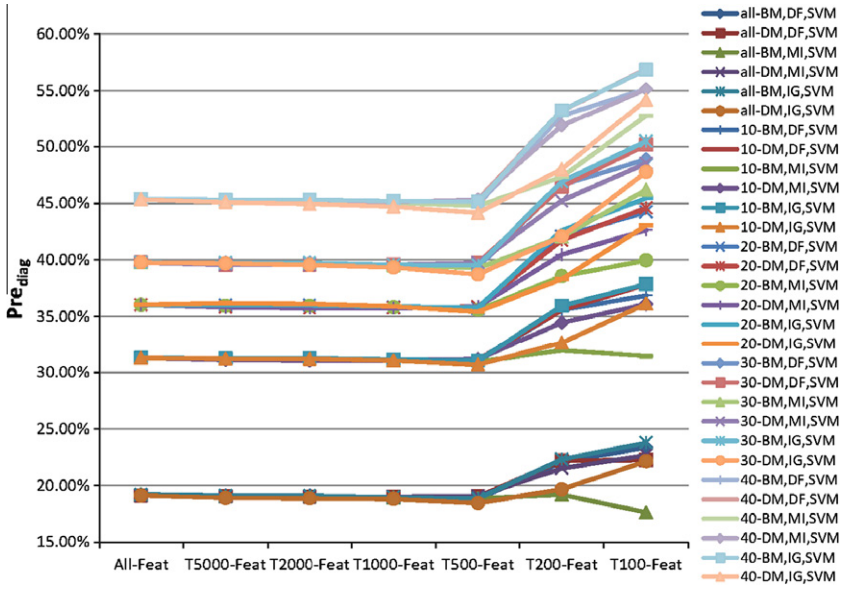

Fig. 9. Results of Pre_{diag} obtained by SVM classifier under different situations.

4.3. Evaluation of normalization performance

The normalization results are shown in Table 3, and the introduced two types of normalization approaches (normalizing symptom names generated by dictionary-based method (NGDM) and normalizing symptom names hidden in the bigram lists generated by bigram-based method (NGBM)) are compared through Pre_{norm}, Rec_{norm} and FM_{norm}.

The F-Measures of NGDM and NGBM are fairly acceptable. NGDM could obtain higher precision than NGBM due to its more accurate symptom name recognition ability. Meanwhile, higher recall is obtained by NGBM benefiting from its extensive and subtle coverage of symptom names. Nevertheless, using NGDM would result in huge amount of repeated labor. Therefore, NGBM is recommended. It does not depend on the complementary symptom name dictionary; however, its performance needs to be further improved.

4.4. Evaluation of automatic diagnosis

The automatic diagnosis results (Pre_{diag}, Rec_{diag} and FM_{diag}) of the proposed framework are represented in Figs. 4–6 under different experimental settings. The highest FM_{diag} (39.15%) could be obtained by "DM, DF, SVM". In these figures, it reveals that SVM could achieve better results than NB. Moreover, SVM could get much better Pre_{diag} and FM_{diag} in most cases when less features are used, although its Rec_{diag} is slightly lower. It also reflects that global feature selection is considered unsuitable for NB in automatic diagnosis, and different feature selection strategies should be investigated or constructed when different types of classifiers are used.

In fact, in clinical practice, several types of disease in raw TCM FCRs only have a small number of referable clinical records and the diagnosis results of several inputted transcripts may not appear in raw TCM FCRs. This phenomenon is one of the reasons results in performance depreciation of automatic diagnosis (see Fig. 7).

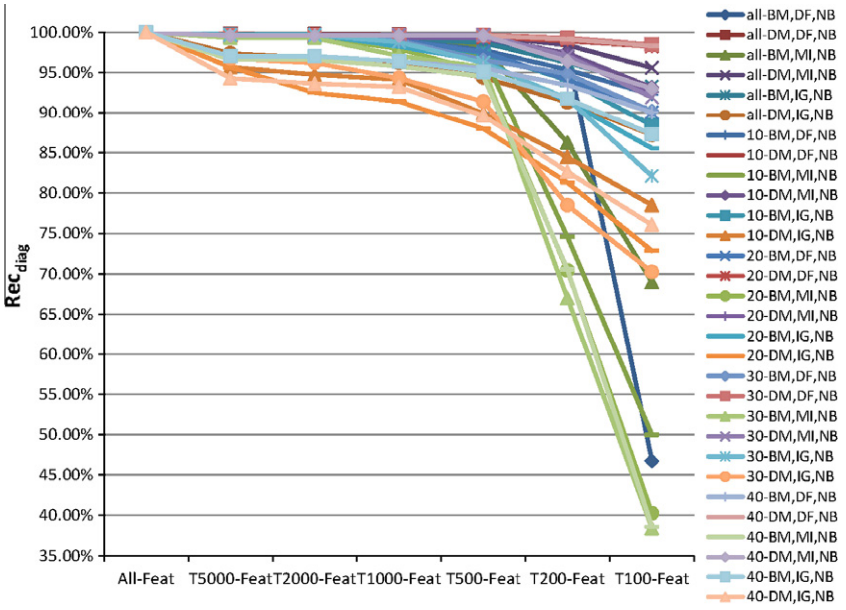

Fig. 10. Results of Rec_{diag} obtained by NB classifier under different situations.

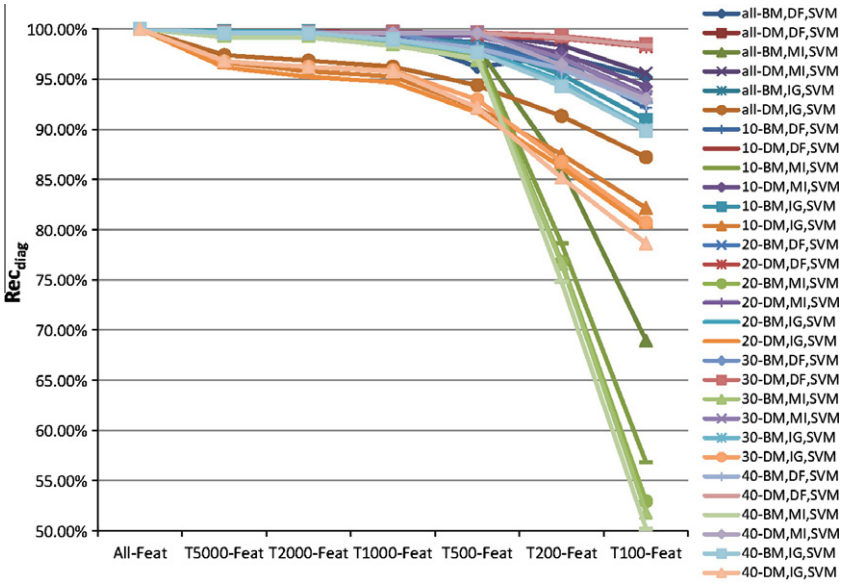

Fig. 11. Results of Rec_{diag} obtained by SVM classifier under different situations.

Therefore, as a further research, a more appropriate automatic diagnosis model should be designed in order to cope with this problem.

Through evaluating the automatic diagnosis results on diverse experimental condition, the worth, necessity and significance of FCRs analysis module, normalization module and feature selection module in the proposed framework are appraised as follows.

The results of the automatic diagnosis by taking different types of feature selection methods without the normalization process are shown in Figs. 8–13. It reveals that the results are significantly improved through using SVM classifier. Only when the number of appearing times of all syndrome labels in CRD is higher than 40 and an appropriate feature selection method (e.g. DF or IG) is used, the performance of NB classifier for automatic diagnosis could begin to be improved. It also shows that although the results of Rec_{diag} dropped slightly in most cases when use less features to diagnose, the results of Pre_{diag} and FM_{diag} by NB classifier have not reduced too much in most cases, even increased a lot by SVM. The results mirror that the clinical records with low frequency syndrome labels would affect the performance of automatic diagnosis. In other words, more knowledge of one disease we have, better performance of automatic diagnosis would be achieved.

In order to illustrate the necessity of the normalization module, the feature sets All-Feat and 100-Feat are normalized by the methods introduced in Section 3.2 (NB classifier and SVM classifier could achieve best FM_{diag} by utilizing these feature sets) and NB classifier and SVM classifier are re-employed. The automatic diagnosis results are listed in Table 4. It is obvious that more than half of the results after normalizing procedure are better. And most of the highest FM_{diag} in each experiment group are achieved by NB and SVM classifiers when the symptom names are normalized. Moreover, not only the performance but also the efficiency of

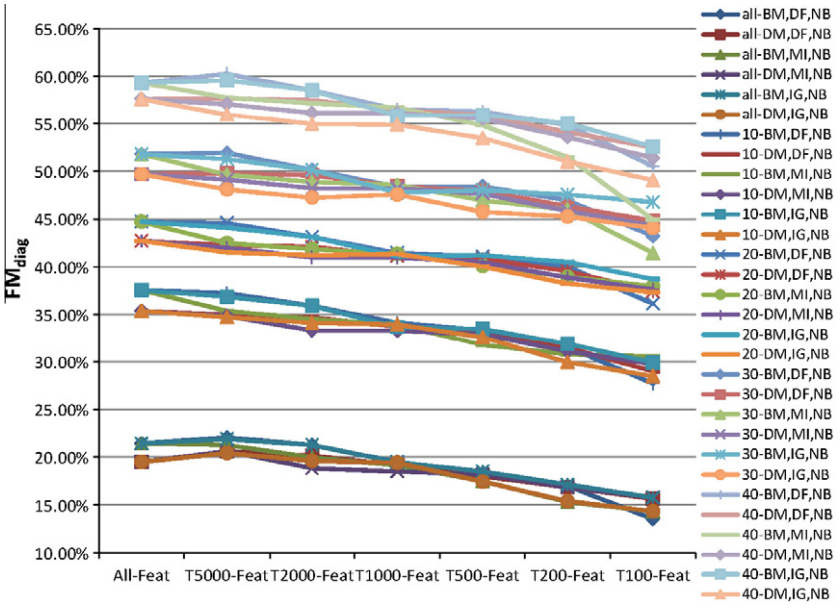

Fig. 12. Results of FM_{diag} obtained by NB classifier under different situations.

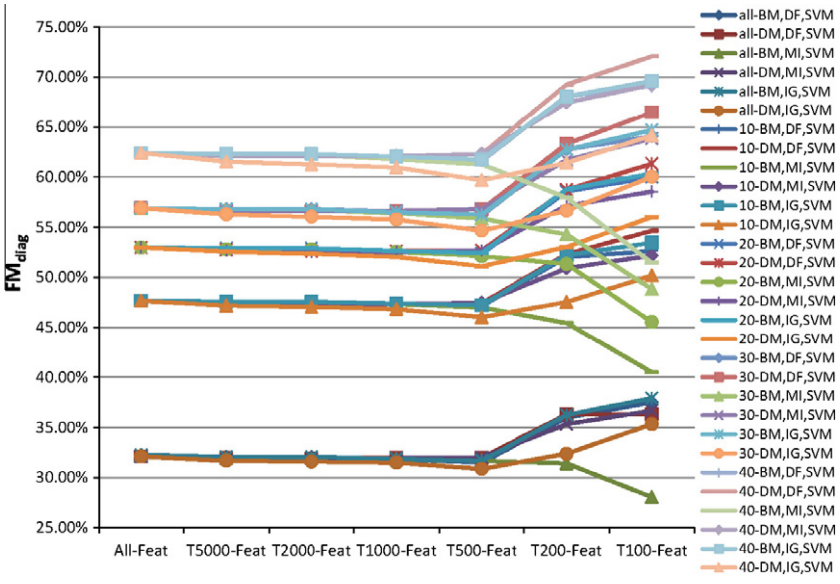

Fig. 13. Results of FM_{diag} obtained by SVM classifier under different situations.

Table 4
Results of the FM_{diag} of two types of automatic diagnosis methods using different feature selection methods with and without symptom name normalization procedure. All-Feat-Norm and 100-Feat-Norm represent the features used to diagnose are normalized by normalization methods (i.e. BM or DM) described before. The bold values are the highest results in each experiment group.

	NB		SVM	
	All-Feat-Norm (%)	All-Feat (%)	100-Feat-Norm (%)	100-Feat (%)
All-BM,DF	21.43	**22.54**	36.08	37.50
All-DM,DF	19.49	21.27	**39.14**	36.31
All-BM,MI	21.43	**22.54**	28.14	28.05
All-DM,MI	19.49	21.27	35.24	36.66
All-BM,IG	21.43	**22.54**	35.01	37.90
All-DM,IG	19.49	21.27	35.07	35.33
10-BM,DF	**40.00**	37.53	53.83	52.63
10-DM,DF	34.57	35.39	**55.45**	54.64
10-BM,MI	**40.00**	37.53	42.77	40.48
10-DM,MI	34.57	35.39	52.34	52.18
10-BM,IG	**40.00**	37.53	52.82	53.45
10-DM,IG	34.57	35.39	50.21	50.18
20-BM,DF	**49.06**	44.72	60.39	59.98
20-DM,DF	42.02	42.64	**62.44**	61.33
20-BM,MI	**49.06**	44.72	46.37	45.53
20-DM,MI	42.02	42.64	60.20	58.55
20-BM,IG	**49.06**	44.72	58.64	60.34
20-DM,IG	42.02	42.64	56.28	56.01
30-BM,DF	**59.06**	51.96	63.57	64.07
30-DM,DF	49.28	50.07	**67.91**	66.47
30-BM,MI	**59.06**	51.96	49.17	48.83
30-DM,MI	49.28	50.07	65.73	63.84
30-BM,IG	**59.06**	51.96	61.52	64.69
30-DM,IG	49.28	50.07	61.02	60.01
40-BM,DF	**63.48**	59.55	67.10	69.12
40-DM,DF	56.94	58.06	**73.32**	72.08
40-BM,MI	**63.48**	59.55	50.74	51.44
40-DM,MI	56.94	58.06	70.86	69.23
40-BM,IG	**63.48**	59.55	65.18	69.62
40-DM,IG	56.94	58.06	65.12	64.14

automatic diagnosis could be improved due to effective feature dimension reduction after symptom name normalization (average feature amount reduction ratio is about 8.52%). In conclusion, it should be emphasized that the normalization module in the framework of automatic diagnosis of TCM utilizing raw FCRs is necessarily included.

5. Discussion

Although Pre_{norm} and FM_{norm} of NGBM in Table 2 is not high enough, it should be emphasized once more that the processes of segmenting the raw TCM FCRs into bigrams and then normalizing the symptom names hidden in the generated bigram lists do not need the extra support of a complementary symptom name dictionary. In other words, it could cut down huge amount of repeated for TCM practitioners and clinical researchers to construct and maintain a complementary symptom name dictionary. NGBM provides an instructional, valuable, and semi-supervised-like approach to extract symptom names from raw TCM FCRs. However, it is necessary to find more advanced NGBM-like methods in order to improve the accuracy of symptom name recognition and normalization results.

The proposed framework is very important and significant for practical aided diagnosis. It provides not only a feasible approach to help TCM researchers to automatically and effectively get scientific hypotheses and clinical diagnosis guidance from raw TCM FCRs directly, but also a referable way for TCM researchers that how to process and utilize raw TCM FCRs automatically. The automatic diagnosis results of the proposed framework have not competed against the other researchers'; however their results are achieved based on the manually well-structured datasets. Their work could not be directly and effectively applied to clinical practice due to the big difference between the well-structured datasets and the raw free-text clinical records. The listed results in Section 4 show that if there are sufficient raw TCM FCRs (i.e. each category of syndrome labels in raw TCM FCRs has sufficient number of clinical records (e.g. more than 40)), the proposed framework could achieve reasonable and acceptable performance in clinical practice.

Although the methods investigated in the proposed framework are traditional, the feasibility and effectiveness of the framework have been verified. With the rapid development of Natural Language Processing, Data Mining and Machine Learning methods, a volume of state-of-the-art methods have been developed, such as Chinese word segmentation and named entity recognition methods for structuring free-text and identifying named entities [23], feature selection method for text classification [24], and multi-class classification learning methods to automatic diagnosis [25,26], etc. These existing methods could be applied in our framework with minor alteration through taking the characteristics of raw TCM FCRs into account. At the same time, the study of diagnosis mechanism should be included in the automatic diagnosis procedure while better performance and higher practical value are aspired.

According to Fig. 3, we could find that there are still a lot of difficulties exist and should be solved, such as how to diagnose more accurately when the syndrome labels are little in the knowledge dataset (i.e. how to take account of the rare disease), how to uninterruptedly learn diagnosis knowledge during a long time accumulating new clinical records in TCM clinical practice, etc. These problems bring not only huge challenges to TCM research but also opportunities to speed up the pace of traditional Chinese medicine modernization.

6. Conclusions

Automatic diagnosis of TCM utilizing raw free-text clinical records is an essential and vital task for applying TCM expert systems

in clinical practice. This problem is attempted for the first time by this paper. A novel framework to tackle this problem is proposed, and a series of appropriate methods are investigated for each module in the framework. At the same time several challenges of the framework are coped with. At last, the experimental results have demonstrated the effectiveness and feasibility of the framework. Through detailed analysis, several remarkable phenomena and problems are pointed out waiting to be further solved.

Fundings

Supported by Provincial Science and Technology Foundation of Sichuan Province (Grant No. 2008SZ0049), National Natural Science Foundation of China (Grant No. 61173182 and 61179071), Specialized Research Fund for the Doctoral Program (Grant No. 20090181110052) and New Century Excellent Talents Fund (Grant No. NCET-08-0370).

Acknowledgments

The authors are grateful to the editor's and the reviewers' comments that help us to improve the quality and merit of this paper. We would like to thank M.S. Xuehong Zhang and M.S. Shengrong Zou for their helpful suggestions to this work and their valuable work on manually structuring the clinical records for us. The authors are also pleased to acknowledge Ms. Fang Yu and Mr. James Chang for their helpful paper revising.

References

[1] Pal SK. Complementary and alternative medicine: an overview. Curr Sci 2002;82(5):518–24.
[2] Barnes PM, Powell-Griner E, McFann K, Nahin RL. Complementary and alternative medicine use among adults: United States, 2002. Semin Integr Med 2004;2(2):54–71.
[3] Molassiotis M, Fernadez-Ortega P, Pud D, Ozden G, Scott JA, Panteli V, et al. Use of complementary and alternative medicine in cancer patients: a European survey. Ann Oncol 2005;16:655–63.
[4] 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 2006;38:219–36.
[5] Lukman S, He Y, Hui SC. Computational methods for traditional Chinese medicine: a survey. Comput Methods Programs Biomed 2007;88:283–94.
[6] Zhou X, Peng Y, Liu B. Text mining for traditional Chinese medical knowledge discovery: a survey. J Biomed Inform 2010;43:650–60.
[7] Wang X, Qu H, Liu P, Cheng Y. A self-learning expert system for diagnosis in traditional Chinese medicine. Expert Syst Appl 2004;26:557–66.
[8] Huang M, Chen M. Integrated design of the intelligent web-based Chinese medical system (CMDS)-systematic development for digestive health. Expert Syst Appl 2007;32:658–73.
[9] Zhang NL, Yuan S, Wang Y. Latent tree models and diagnosis in traditional Chinese medicine. Artif Intell Med 2008;42:229–45.
[10] Wang Y, Yu Z, Jiang Y, Xu K, Chen X. Automatic symptom name normalization in clinical records of traditional Chinese medicine. BMC Bioinform 2010;11:40.
[11] Maciocia G. The foundations of Chinese medicine: a comprehensive text for acupuncturists and herbalists. Churchill Livingstone; 2005.
[12] Norman J. Chinese. United Kingdom: Cambridge University Press; 1988.
[13] Yang X, Liang Z, Luo Y, Yin J. A classification algorithm for TCM syndromes based on P-SVM. In: Proceedings of 2005 international conference on machine learning and cybernetics, vol. 6, Guangzhou, China; 2005. p. 3692–7.
[14] Kulick S, Bies A, Liberman M, Mandel M, McDonald R, Palmer M, et al. Integrated annotation for biomedical information extraction. In: Proceedings of human language technology conference and north american chapter of the association for computational linguistics annual meeting workshop: Biolink; 2004. p. 61–8.
[15] Tsuruoka Y, Tsujii J. Bidirectional inference with the easiest-first strategy for tagging sequence data. In: Proceedings of joint conference on human language technology and empirical methods in natural language processing. Vancouver, B.C., Canada; 2005. p. 467–74.
[16] Tsuruoka Y, Tateishi Y, Kim JD, Ohta T, McNaught J, Ananiadou S, et al. Developing a robust Part-of-Speech tagger for biomedical text. In: Proceedings of the 10th Panhellenic conference on informatics, LNCS 3746; 2005. p. 382–92.
[17] Nie J-J, Gao J, Zhang J, Zhou M. On the use of words and *n*-grams for Chinese information retrieval. In: Proceedings of the fifth international workshop on information retrieval with Asian languages, IRAL2000, Hong Kong; 2000.
[18] Yang Y, Pedersen JP. A comparative study on feature selection in text categorization. In: Proceedings of the fourteenth international conference on machine learning; 1997. p. 412–20.
[19] Kononenko I. Inductive and bayesian learning in medical diagnosis. Appl Artif Intell 1993;7(4):317–37.
[20] Ramana BV, Babu MSP, Venkateswarlu NB. A critical study of selected classification algorithms for liver disease diagnosis. Int J Database Manage Syst 2011;3(2):101–14.
[21] Manning CD, Schutze H. Foundations of statistical natural language processing. Cambridge, Massachusetts: The MIT Press; 1999.
[22] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. <http://www.csie.ntu.edu.tw/≈cjlin/papers/libsvm.pdf>; 2001.
[23] Gao J, Li M, Wu A, Huang C-N. Chinese word segmentation and named entity recognition: a pragmatic approach. Comput Linguist 2005;31(4).
[24] Singh SR, Murthy HA, Consalves TA. Feature selection for text classification based on gini coefficient of inequality. In: Proceedings of the fourth international workshop on feature selection in data mining, vol. 10; 2010. p. 76–85.
[25] Li T, Zhang C, Ogihara M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics 2004;20(15):2429–37.
[26] Aly M. Survey on multiclass classification methods. <http://www.vision.caltech.edu/malaa/publications/aly05multiclass.pdf>; 2005.