World Wide Web (2022) 25:1447–1468
https://doi.org/10.1007/s11280-021-00981-5

Bayesian networks and chained classifiers based on SVM for traditional chinese medical prescription generation

Yingpei Wu[1,6] · Chaohan Pei[1,2] · Chunyang Ruan[3] · Ruofei Wang[2] · Yun Yang[4,5] · Yanchun Zhang[2,6]

Received: 16 March 2021 / Revised: 28 September 2021 / Accepted: 22 November 2021 /
Published online: 8 December 2021

Abstract

Traditional Chinese Medicine(TCM) is playing an increasingly prominent role in lung cancer treatment, as it can prolong patients' survival, improve their quality of life, and reduce the adverse effects of radiotherapy and chemotherapy. However, the effectiveness of TCM treatment depends more on the personal experience of doctors, and the standardization of TCM prescriptions needs to be strengthened. In this study, we use TCM clinical prescriptions to train a standardized TCM prescription generation model to provide an auxiliary prescription reference for physicians. However, in our initial experiments, we found two severe problems in the dataset. The first problem is a strong correlation between each herb; for instance, some herbs often appear together to treat specific symptoms. The second is a severe class imbalance within each label, a few herbs always appear in most prescriptions, but most herbs have a low frequency of occurrence in the total dataset. To solve the correlation between each herb label, we adopt the Bayes Classifier Chain(BCC) algorithm, whose basic classifier is Cost-Sensitive SVM targeted to the class imbalance of the label. Based on this, we also improve the BCC method according to the characteristics of TCM prescription dataset. In our BCC classifier, the Directed Acyclic Graph (DAG) construction method has high interpretability in the scenario of TCM prescription. After combining multi-label learning algorithms with several SVM algorithms and comparing their performance in detail, we find that BBC+CS-SVM best deals with class imbalance within the label in multi-label classification problems.

✉ Chunyang Ruan
cyruan16@fudan.edu.cn

[1] School of Computer Scinece, Fudan University, Shanghai, China

[2] Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, China

[3] School of Economics and Finance, Shanghai International Studies University, Shanghai, China

[4] Department of Oncology, Shanghai Municipal Hospital of Traditional Chinese Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai 200071, China

[5] Oncology Institute of Traditional Chinese Medicine, Shanghai Municipal Hospital of Traditional Chinese Medicine, Shanghai 200071, China

[6] Oujiang laboratory, (Zhejiang lab for Regenerate Medicine, Vision and Brain Health), Wenzhou, Zhejiang 325001, China

Keywords Multi-label classification · Bayes classifier chain · Cost sensitive SVM · TCM

1 Introduction

Lung cancer is one of the most common malignant tumors today and has become the leading cause of human deaths due to cancer [21]. Many clinical and experimental studies have shown that TCM combined with tumor radiation therapy and chemotherapy can alleviate side effects and improve patients' quality of life [10]. It can be seen that in the comprehensive treatment of lung cancer, TCM has become one of the critical components.

However, unlike the theoretical system and treatment methods of modern Western medicine, TCM emphasizes personalized diagnosis and treatment, and the therapeutic efficacy is highly related to the clinical experience of the doctor. For example, the prescriptions made up by different TCM physicians, which contain a group of herbs, may differ significantly for the same patient. Therefore, through a comprehensive analysis of the clinical prescription, the knowledge and rules implied in the prescription are found. It plays a vital role in the modernization and standardization of diagnosis and treatment of TCM. In related studies, most prescriptions are from the classic literature of TCM. However, these prescriptions are too old and simple to meet the personalized demand of modern medicine. Fortunately, we obtained more than 10000 clinical prescriptions of TCM for lung cancer from our cooperative hospital and employed these data for our experiments.

As shown in Table 1, the clinical TCM electronic medical record is mainly composed of two parts (i.e., symptoms and herbs). The set of symptoms observed by the doctor is depicted in the first row. Based on these observed symptoms, a group of herbs prescribed by the doctor are shown in the second row.

In order to construct an interpretable and standardized prescribing process for practitioners' reference, we propose a novel multi-label classifier, which takes in a group of symptoms and outputs a set of herbs.

In our initial experiments, we encountered two severe difficulties in the data. The first is a strong correlation between each herb; for instance, some herbs often appear together to treat specific symptoms. The second is a very serious class imbalance within each label, a few herbs always appear in most prescriptions, but most herbs have a low frequency of

Table 1 A TCM clinical prescription of lung cancer

Sympotms	Tongue reddish(舌淡红), Deficiency of both qi and yin(气阴两虚), Thin tongue fur(苔薄), Pulmonary malignant tumor(肺恶性肿瘤), Pulse fine(脉细)
Herbs	Desert cistanche(苁蓉) , Uncooked rice kernels(生米仁) , Astragali radix(生黄芪), Selaginella doederleinii Hieron(石上柏) , Asparagus fern(天冬), Hedyotis diffusa(蛇舌草), Lossy privet fruit(女贞子), Raw atractylodes(生白术), Salvia chinensis(石见穿), Edible tulip(山慈菇), Herba epimedii(仙灵脾), Akebia fruit(预知子), Radix glehniae(北沙参), Gizzard pepsin(鸡内金)

occurrence in the total dataset. In our whole dataset, 357 labels are all the herbs in our dataset. Each label is similar to each item in the herbs row of Table 1. However, there are 255 herb labels, and their positive samples account for less than 3.3% of the total samples. In other words, 255 herbs only appear in 3.3% of prescriptions. Figure 1 shows the imbalance of the herbs in our dataset.

According to the presumption that the labels are independent of each other, the BR algorithm decomposes the multi-label classification task into a set of separate binary classification tasks. Due to the complexity of reality that many correlations are between different labels, the BR algorithm is difficult to meet the requirements of actual multi-classification tasks.

To solve the challenge of label correlations in the dataset, Label Power-set (LP) method [22] converts the multi-label classification into a multi-class problem by training a classifier on the label combinations in the training dataset. However, with the number of escalated labels, the computational complexity of LP increases exponentially. Based on BR, the Classifier Chain model (CC) adopts several binary classifiers and transforms the multi-label learning task into a set of ordered binary classification problems. The input of each binary classifier is based on the prediction results of previous classifiers. However, the major deficiency of the CC model is how to effectively analyze the correlations between labels and determine the order of classifiers.

Taking a Directed Acyclic Graph (DAG) that models the dependence relationship of the labels, L Enrique et al. [20] proposed Bayesian Classifier Chains (BCC). BCC trains each classifier from the root node and then delivers the results of the parent classifier to the child classifier. Inspired by the BBC method, we construct a DAG for the herb label set based on the specific attributes in the TCM prescriptions to meet the challenge of label correlations.

The solutions to solve class imbalance can be broadly divided into two categories: changing the data sampling strategy and cost-sensitive learning methods. Changing the data sampling strategy is to modify the sample's distribution by adopting data resampling methods, including oversampling, undersampling, and synthetic sampling [1, 3,

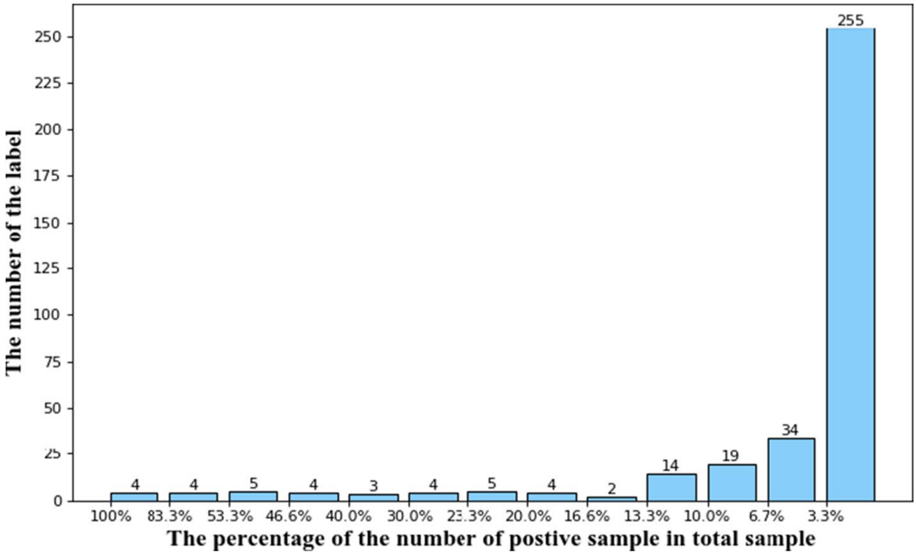


Fig. 1 Analysis on the imbalance of herbs

7]. In our initial experiments, however, the performance after changing the data sampling strategy was unsatisfactory due to the extremely high false positive rate. Therefore, in this paper we mainly investigate the cost-sensitive learning methods. Inspired by the success of Cost-Sensitive SVM(CS-SVM) to deal with class-imbalanced problem [13], We adopt the CS-SVM as the primary classifier for the Classifier Chain model. The proposed model can not only solve the class imbalance problem, but also implement the cost-sensitive Bayes decision rule and make the model risk approximate the cost-sensitive Bayes risk. The contributions of our work are summarised as follows:

- A multi-label classification algorithm based on a Bayesian classifier chain is proposed for the label correlation problem. The algorithm simulates the prescription thinking of TCM practitioners and is used for the TCM prescription generation task.
- We propose a CS-SVM to solve the class imbalance problem in prescriptions. We illustrate the derivation process of the CS-SVM and theoretically demonstrate that it can better solve the class imbalance problem.
- We carry out extensive experiments to compare the proposed method and several combinations of multi-label classification and basic classifier algorithms and two deep learning-based methods. The results show that the performance of the proposed model is better than the others. Furthermore, a case study is conducted, and the results show that our proposed model has good predictive capability for rare herbs and has high clinical application value in TCM.

The rest of the paper is organized as follows. In Section 2, we introduce current methods in TCM Knowledge Discovery. The proposed model for TCM prescription generation is presented in Section 3. The experiment results and a case study are presented in Section 4. Finally, we conclude this paper and discuss the future work in Section 5.

2 Related work

2.1 TCM knowledge discovery

With the continuous improvement of artificial intelligence and data mining technology, more and more attention has been paid to mining TCM knowledge from prescriptions. The topic model considers TCM prescriptions as documents where the herbs and symptoms are words. Ma et al. [12] presented a “syndrome-symptom” model to extract the relation among the topics of symptom and syndrome. Yao et al. [26] proposed a topic model which depicts the generative process of prescriptions in TCM theories and further includes domain knowledge into the topic model. Recently graph representation learning-based TCM knowledge discovery has been a hot research topic. Ruan et al. [17–19] modeled TCM prescriptions as graphs where herbs and symptoms are nodes and the co-occurrence relationship as the edge to mine complex relation between herbs and symptoms. With the deep learning techniques constantly developing, some researchers have adopted deep learning techniques to prescription mining. Li et al. [9] proposed an attention based Seq2Seq model to automatically generate prescriptions. Li et al. [8] utilized a transformer based Seq2Seq model to imitate the prescribing process and mine the rules of TCM prescription. Taking chronic obstructive pulmonary disease as the background, Xu et al. [25] studied TCM syndrome differentiation based on neural network.

2.2 Classifier chain

Jesse et al. [16] proposed the CC model to tackle the multi-label classification task, which can take label dependencies into account and achieves the computational efficiency of the binary relevance method. Following the basic CC method, many improvements have emerged. The Probabilistic Chain Classifier (PCC) algorithm was proposed by Dembczynski et al. [5], which is primarily used in the probabilistic framework of CC. While PCC can better take into account the correlation between labels, its time complexity is often unacceptable. Because of the label ordering having a dramatic effect on the performance of prediction, Gonçalves et al. [6] proposed a Genetic Algorithm for ordering Classifier Chains (GACC) algorithm to optimize the label order in classifier chains.

2.3 Cost sensitive SVM

Based on statistical learning theory, SVMs have been widely employed in pattern recognition and classification tasks. There are two types of cost-sensitive SVMs used to address the class imbalance problem. The first one is called the Biased Penalties SVM (BP-SVM) [2, 24], and its mechanism is to apply two penalty factors P_1 and P_{-1} to the positive and negative slack variables of SVM in the training process. It is realized by converting the original SVM problem to

$$\begin{aligned} & arg \min_{w,b,\zeta} \frac{1}{2}||w||^2 + P\left[P_1 \sum_{\{i|y_i=1\}} \zeta_i + P_{-1} \sum_{\{i|y_i=-1\}} \zeta_i\right] \\ & \text{s.t. } y_i(w^T x + b) \geq 1 - \zeta_i \end{aligned} \tag{1}$$

BP-SVM is subjected to an apparent defect; that is when the training data is separable, its ability to implement cost-sensitive strategy is limited. During parameter optimization, the model does not refine the penalty parameter P_1 and P_{-1}, but selects the larger slack penalty P, and then makes the slack variable ζ zero-valued. The optimization of BP-SVM is transformed into the standard support vector machine, and the separating hyperplane is placed in the middle of two classes (instead of specifying a larger margin for one of them). The second one is the CS-SVM [13], which optimized the hinge loss function by a cost-sensitive method instead of relying on penalty terms. We will elaborate on it in the following section.

3 Methodology

In this paper, We consider the prescription generation task as a multi-label classification problem. In the following, we use uppercase boldface letters to denote vectors and the normal lowercase letter to a scalar or a component of a vector. Each training sample $(\mathbf{X}_i, \mathbf{Y}_i)$ consists of a symptom set and a herb set, where $\mathbf{X}_i$ and $\mathbf{Y}_i$ denote a vector of symptoms and herbs respectively. For each $\mathbf{X}_i = [x_1, x_2, \cdots, x_S] \in \{-1, 1\}^S$,

$\mathbf{Y}_i = [y_1, y_2, \cdots, y_T] \in \{-1, 1\}^T$, the S and T are the dimensions of the input and output vector respectively. In the prescription prediction scenario, S and T represent the number of symptoms and herbs. The jth component of the vector $\mathbf{X}_i$ is 1 if the symptom set of a prescription contains the symptom $s_j, (j = 1, \cdots, M)$, otherwise it is -1, and it is the same with the herb vector $\mathbf{Y}_i$.

Our goal is to train a multi-label classifier $F(\cdot)$ satisfying the functional relationship $\mathbf{Y} = F(\mathbf{X})$ on the TCM clinical prescriptions.

3.1 BCC algorithm

The framework of the BCC algorithm is illustrated in Figure 2 and Algorithm 1, which consists of 2 parts:

1. Building the order of classifiers in the chain and constructing the Directed Acyclic Graph(DAG),
2. Optimizing the BCC classifier based on the DAG, which is described in Algorithm 3.

3.1.1 Construct the directed acyclic graph (DAG)

In general, building a Bayesian network is an NP-hard problem, but in our prescription prediction task, we can optimize this process by analyzing the dataset features. First, we count the frequency of each herb in more than 10,000 prescriptions and then rank all herb labels according to the number of occurrences. We find that the higher the occurrence frequency of the herbs, the higher their importance. Doctors tend to give priority to these herbs when prescribing.

In the TCM diagnosis scenario, TCM doctors usually consider a basic formula including commonly used herbs first, and then judge whether to use rare herbs on the basic formula. Therefore, we can determine the order of the labels in the classifier chain, i.e., the direction of training the classifier is from herbs with a high frequency of occurrence to low frequency. In label sample matrix $H \in \{-1, +1\}^{N \times T}$, where N is the number of samples, T is the number of herb labels which is 357, and the order of the matrix column vectors is in descending order of herb frequency.

Secondly, with the label sample matrix H, we calculate the Pearson correlation coefficient for each herb and then construct the Pearson correlation coefficient matrix $P(LxL)$. After consulting with the physicians, we decided to set the correlation coefficient threshold to 0.2. In other words, if $|p_{i,j}| > 0.2$, we consider that there is a correlation between herb i and herb j, and then the value of $p_{i,j}$ is 1, otherwise it is 0. After discretizing the correlation matrix, the matrix $P(LxL)$ is composed of 0 or 1, which naturally forms an adjacency matrix and represents a graph (i.e., G = <V, E>). Next, we use the graph G to construct a DAG.

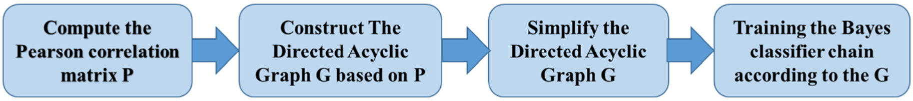

Compute the Pearson correlation matrix P
Construct The Directed Acyclic Graph G based on P
Simplify the Directed Acyclic Graph G
Training the Bayes classifier chain according to the G

Fig. 2 The framework of BCC

Algorithm 1 The framework of BCC training process.

```
Input: The symptom features matrix $S \in \{-1,+1\}^{N \times M}$. The label sample matrix $H \in \{-1,+1\}^{N \times T}$.;
Output: The BCC classifier $F$;
Based on label sample matrix $H$, Construct the Pearson correlation coefficient matrix $P \in \{-1,+1\}^{L \times L}$.
Define the threshold $t$, $\forall p_{i,j} \in P$, if $|p_{i,j}| > t$, $p_{i,j} \leftarrow 1$ else $p_{i,j} \leftarrow 0$. Then obtain the adjacent matrix $G$ based on $P$.
Apply the Algorithm 2, input the DAG adjacent matrix $G$, get simplified DAG adjacent matrix $G'$.
Use the DAG: $G' =< V, E' >$ as the classifier chain order of BCC algorithm, then call Algorithm 3.
```

Algorithm 2 Simplification of the DAG.

```
Input: The adjacent matrix $D$ of the original DAG: $G =< V, E >$,
       Sign array $Signlist = \{0\}^{num(V)}$, where $num(V)$ is the vertex number of $V$,
       The different connected component flag, $Sign = 0$,
       The matrix $D' = \{0\}^{num(V) \times num(V)}$.
Output: The adjacent matrix $D'$ of the simplified DAG: $G' =< V, E' >$;
function DFS$(i, num(V))$ :
   For $j = i + 1$; $j < num(V)$; $j++$ do
      if $D[i][j] = 1$ and $Signlist[j]! = Sign$ then
         $D'[i][j] \leftarrow 1$ //        Add directed edge $e =< i, j >$ to $E'$
         $Signlist[j] \leftarrow Sign$
         DFS$(j, num(V))$
      end if
   return
end function

For $i = 0$; $i < num(V)$; $i++$ do
   $Sign \leftarrow Sign + 1$
   DFS$(i, num(V))$
retrun $D'$
```

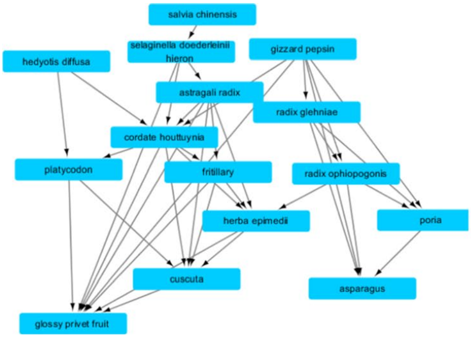
Algorithm 3 Training BCC based on DAG.

```
Input: The adjacent matrix $D'$ of the simplified DAG: $G' =< V, E' >$;
       The array of sum about every node in-degree, $Sum\_in$;
       The additional feature sets of all herb label nodes, $T_0, T_1, T_2, \cdots = \emptyset$;
       The basic classifier $f(\cdot)$;
       The symptom feature set $X$.
Output: The BCC classifier $F$;
for $k = 0$;$k < num(V)$;$k++$ do
    $Sumin[k] \leftarrow sum(D[\cdot][k])$// compute the indegree of each node

function Training$(i, num(V), T')$ :
    $Sum\_in[i] \leftarrow Sum\_in[i] - 1$;
    if $Sum\_in[i]! = 0$ then
        return
    else
        The eventual input set $X_i'$ for $f_i(\cdot)$: $X' \leftarrow X \cup T_i$
        Use the sample in input feature set $X'$ and target set $y_i$ train $f_i(\cdot)$
        for $j = i + 1$; $j < num(V)$; $j++$ do
            $T_i \leftarrow T_i \cup T'$;
            Training$(j, num(V), T_i)$
        return
    end if
end function

do
    for $u = 0$;$u < num(V)$;$u++$ do
        if $Sum\_in[u] = 0$ then
            Training$(u, num(V), T_u)$
            break
        end if
while $u! = num(V) - 1$
return BCC classifier $F(\cdot) = \left[f_0, f_1, f_2, \ldots, f_{num(v)-1}\right]$
```

Thirdly, we build a DAG $G =< V, E >$ consisting of a node set V, a link set E. Each



(a) Original DAG (b) The simplified DAG by DFS pruning

Fig. 3 A comparison between the original and simplified DAG

node v_i corresponds to a herb label y_i. We suppose that if $p_{i,j} = 1$ and $i > j$, then the directed edge $< y_i, y_j > \in E$. However, the DAG suffer from a handicap that if there exists a path $r_{i,j}$ from i to j, we find that there will be a large number of directed edges $< v_k, v_j >$, where $\{v_k | v_k \in r_{i,j}\}$. To overcome this, we utilize the Deep First Search (DFS) algorithm to prune the useless edge, which is elaborated in Algorithm 2 and Figure 3.

3.2 Cost sensitve SVM

3.2.1 Bayesian consistency of standard binary classifier

The goal of binary classification is to train a function $h : x \rightarrow 0, 1$ called a classifier to predict Y given X using the training dataset. From a statistical perspective, the feature vectors of properties $\mathbf{x}$ and class labels y can be considered as random variables with probability distributions $P_X(\mathbf{x})$ and $P_Y(y)$ respectively. We define the classifier function as $h(\mathbf{x}) = sign[p(\mathbf{x})]$, where the function $p : \mathcal{X} \rightarrow \mathbb{R}$. The non-negative loss function for each $(p(\mathbf{x}), y)$ pair is $L(h(\mathbf{x}), y)$. We achieve the goal of minimizing the conditional risk by minimizing the expected loss

$$\begin{aligned} E_{Y|\mathbf{X}}[L(p(\mathbf{x}), y)|\mathbf{X} = \mathbf{x}] =& P_{Y|\mathbf{X}}(1|\mathbf{x})L(p(\mathbf{x}), 1) \\ & + (1 - P_{Y|\mathbf{X}}(1|\mathbf{x}))L(p(\mathbf{x}), -1). \end{aligned} \tag{2}$$

To better comprehend this formula, we rewrite the function $p(\mathbf{x}) = f(\delta(\mathbf{x}))$, where $\delta(\mathbf{x}) = P_{Y|\mathbf{X}}(1|\mathbf{x})$ is the posterior probability. The Link function is defined as $f : [0, 1] \rightarrow \mathbb{R}$, which made a connection to Bayesian decision rules in this way. The Bayes error rate of the data distribution is the probability that an instance is misclassified by a classifier that knows the true class probabilities given the predictors. We hope to minimize the conditional risk of the model and make it close to the Bayes error rate, so that the model is theoretically optimal. To minimize the conditional risk when the true probability distribution is known and the loss function L is determined, we can choose a suitable link function f. As for how to choose the appropriate link function $f$to make conditional risk (3) approximate Bayesian error rate, this is described in detail in [27].

We extend this minimized problem to the cost-sensitive version. The detailed derivation can refer to the appendix or [13]. Now we give the result directly.

Suppose ϕ is a specific form of the loss function L, for example the hinge loss function in SVM $\phi(yf) = \lfloor 1 - yf \rfloor_+$. In CS-SVM, the loss function ϕ may have different forms in false positive and false negative, so we can define the cost-sensitive loss function in a uniform form

$$\begin{aligned} L_{\phi,C_1,C_{-1}} =& \phi_{C_1,C_{-1}}(yf) \\ =& \begin{cases} \phi_1(f), & \text{if } y = 1; \\ \phi_{-1}(f), & \text{if } y = -1. \end{cases} \end{aligned} \tag{3}$$

We get the cost sensitive conditional risk from (2) and (3)

$$C_{\phi,C_1,C_{-1}}(\eta,f) = \eta\phi_1(f) + (1-\eta)\phi_{-1}(-f), \tag{4}$$

which can be minimized by the link function

$$f^*_{\phi,C_1,C_{-1}}(\eta) = arg \min_f C_{\phi,C_1,C_{-1}}(\eta,f) \tag{5}$$

So we get the minimum conditional risk finally

$$\begin{aligned} C^*_{\phi,C_1,C_{-1}}(\eta) =& \eta\phi_1(f^*_{\phi,c_1,C_{-1}}) \\ &+ (1-\eta)\phi_{-1}(-f^*_{\phi,C_1,C_{-1}}) \end{aligned} \tag{6}$$

3.2.2 Cost sensitive SVM loss function

In this section, the SVM hinge loss function will be updated to the cost-sensitive version.

For the standard SVM, the loss function is the hinge loss $\phi(yf) = \lfloor 1 - yf \rfloor_+$, and the optimal link function is

$$f^*_\phi(\eta) = sign(2\eta - 1) \tag{7}$$

and the minimum conditional risk is

$$\begin{aligned} C^*_\phi(\eta) =& 1 - 2|2\eta - 1| \\ =& \eta\lfloor 1 - sign(2\eta - 1)\rfloor_+ + (1-\eta)\lfloor 1 + sign(2\eta - 1)\rfloor_+ \end{aligned} \tag{8}$$

We extend the optimal link function of the standard SVM to the cost-sensitive setting, and obtain the optimal link function of the CS-SVM

$$f^*_{\phi,C_1,C_{-1}}(\eta) = sign((C_1 + C_{-1})\eta - C_{-1}) \tag{9}$$

The process of proof about the optimality of this link function can refer to [13], in which they give some requirements to judge whether a link function is optimal. Similar to the conditional risk of the standard SVM, we obtain the minimum conditional risk of the CS-SVM

$$\begin{aligned} C^*_{\phi,C_1,C_{-1}}(\eta) =& \eta\lfloor e - d \cdot sign((C_1 + C_{-1})\eta - C_{-1})\rfloor_+ + \\ & (1-\eta)\lfloor b + a \cdot sign((C_1 + C_{-1})\eta - C_{-1})\rfloor_+ \end{aligned} \tag{10}$$

where

$$d \geq e, \quad a \geq b, \quad \frac{C_{-1}}{C_1} = \frac{a+b}{d+e} \tag{11}$$

The function of this condition is guaranteeing the Bayesian consistency, and the details can refer to [13]. When the a, b, d, e are positive, we can easily find that

$$sign((C_{-1}+C_1)\eta - C_{-1}) = \begin{cases} 1, & \text{if } \eta \geq \gamma \\ 0, & \text{if } \eta = \gamma \\ -1, & \text{if } \eta \leq \gamma \end{cases} \tag{12}$$

where $\gamma = \frac{C_{-1}}{C_1 + C_{-1}}$.

If $\eta < \gamma$, the risk is

$$C^*_{\phi,C_1,C_{-1}}(\eta) = \eta\lfloor e+d \rfloor_+ + (1-\eta)\lfloor b-a \rfloor_+ \tag{13}$$

Obviously, we want to minimize the risk, the $\lfloor e+d \rfloor_+ = (e+d) \geq 0$, so we cannot modify it. But if $a \geq b$,the $\lfloor b-a \rfloor_+$ will be 0, otherwise will be a positive number. Therefore, make $a \geq b$ can minimize $C^*_{\phi,C_1,C_{-1}}$, we can get $d \geq e$ in the similar way. Finally, similar to the form of the hinge loss of the standard SVM, we get the loss function of the CS-SVM

$$\phi_{C_1,C_{-1}}(yf) = \begin{cases} \lfloor e - df \rfloor_+, & \text{if } y = 1; \\ \lfloor b + af \rfloor_+, & \text{if } y = -1; \end{cases} \tag{14}$$

The hinge loss function has four degrees of freedom, controlling the margin and slope of each of the two classes. The positive class is divided by margin $\frac{e}{d}$ and slope d of hinge loss, and the negative class is divided by margin $\frac{b}{a}$ and slope a of hinge loss.

3.2.3 Cost sensitive SVM algorithm

Although the hinge loss function of the CS-SVM has four degrees of freedom, the conditional risk function $C^*_{\phi,C_1,C_{-1}}$ has only two degrees of freedom. Therefore, after obtaining the proportional relationship between the margins of the two classes, we assume that the weight of the positive class is more important, which requires the slope and margin of the positive class to be greater than the negative class,

$$\frac{e}{d} \geq \frac{b}{a} \quad d \geq a \tag{15}$$

and then set the $\frac{e}{d} = 1$, $e = d = C_1$ to fix the margin of positive class. Similarly, we also need the ratio between a and b. So setting b = 1, according to (12) we get $a = 2C_{-1} - 1$. After obtaining the values of a,b,c,d and e, we can derive the minimum conditional risk of the cost-sensitive SVM by (10):

$$\begin{aligned} C^*_{\phi,C_1,C_{-1}}(\eta) =& \eta\lfloor C_1 - C_1 \cdot sign((C_1 + C_{-1})\eta - C_{-1}) \rfloor_+ + \\ & (1-\eta)\lfloor 1 + (2C_{-1} - 1) \cdot sign((C_1 + C_{-1})\eta - C_{-1}) \rfloor_+ \end{aligned} \tag{16}$$

with $C_{-1} \geq 1, C_1 \geq 2C_{-1} - 1$ to satisfy the condition of (13). The intuitive explanation is that the positive class has a larger margin, which will shift the separating hyperplane toward the negative class, and then increase the risk of cost in case of misclassification.

We modify the risk of the standard SVM by cost-sensitive learning as follows:

$$\begin{aligned} &arg \min_{w,b} \sum_{\{i|y_i=1\}} \lfloor C_1 - C_1(w^T x_i + b) \rfloor_+ \\ &+ \sum_{\{i|y_i=-1\}} \lfloor 1 + (2C_{-1} - 1)(w^T x_i + b) \rfloor_+ + \mu ||w||^2 \end{aligned} \tag{17}$$

then deduce to a primer optimization problem

$$\begin{aligned} &arg \min_{w,b,\xi} \frac{1}{2} ||w||^2 + C \left[\beta \sum_{\{i|y_i=1\}} \xi_i + \lambda \sum_{\{i|y_i=-1\}} \xi_i \right] \\ &\text{s.t. } y_i(w^T x + b) \geq 1 - \xi_i, y_i = 1 \\ &\quad\; y_i(w^T x + b) \geq \kappa - \xi_i, y_i = -1 \end{aligned} \tag{18}$$

with

$$\beta = C_1 \quad \lambda = 2C_{-1} - 1 \quad \kappa = \frac{1}{2C_{-1} - 1}$$

In the quadratic programming problem, there are three parameters β, γ, κ that determine the cost sensitivity. The β, γ decide the relative weights of margin violations and pay more attention to positive class on the constrain that $C_{-1} \geq 1, C_1 \geq 2C_{-1} - 1$. In the optimization procedure of BP-SVM, the model prefers to increase the general penalty term C rather than to adjust the cost-sensitive parameter C_1, C_{-1}, so the separating hyperplane will be set in the middle of the two types of training data, and BP-SVM will degenerate into standard SVM. The proposed model can achieve cost-sensitive classification not only by changing the slack penalty C, but also by directly decreasing the margin of the majority class to move the separating hyperplane to the majority class. In the next section, we will intuitively explain the distinction between standard SVM, BP-SVM, CS-SVM.

3.2.4 Distinction between standard SVM, BP-SVM, CS-SVM

In this section, we consider the majority and minority class as the negative and positive class, respectively. In the practical situation of clinic prescriptions, the positive class often is a minority class but more important. Figure 4 is an example of imbalanced data. Figure 4(a) is the comparison between standard SVM and BP-SVM. It can be seen that the separating hyperplane is very close to the minority class interior. This is because the number of the outlier dots of the majority class is more than minority class, but the slack variable penalty factor C for these two classes are equivalent, in other words, the model has the same tolerance to these two classes, so if the quantity of certain class is much more than the other, the outlier of the latter may be ignored, such as the right three red dots in Figure 4(a). A fundamental problem is a small number of minority training samples, but the distribution of the minority, in reality, may not like the training dataset; for example, the hollow dots in Figure 4(a) is the possible positive sample, but they are not included in train dataset. This case can be improved by applying BP-SVM.

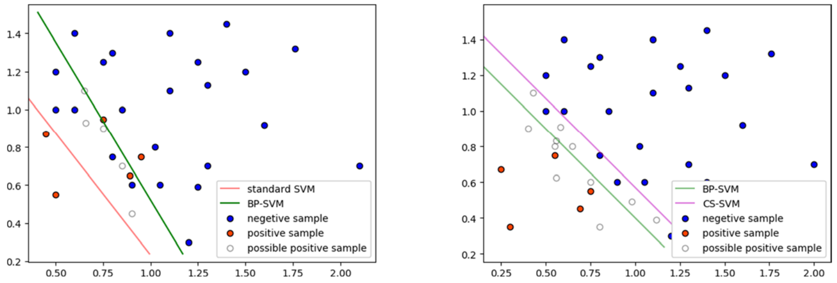
If we do not want to change the distribution of the training dataset, we can alter the penalty factor C of two classes by means of increasing the C_{+1} if we want to pay more attention to positive class, so that the model will be more unbearable to the minority outlier dots even if their quantity is few. The separating hyperplane of BP-SVM is pushed to the majority class, and it can classify the potential minority positive dots correctly.

Although the BP-SVM can perform better than the standard SVM in the situation of data inseparable, in the separable dataset, the BP-SVM will degenerate into standard SVM. In Figure 4(b), it seems that the training sample of two classes can be separated. The BP-SVM has the general penalty C that controls the general tolerance of outlier and the cost-sensitive penalty C_{+1}, C_{-1} can be a proportion to control the tolerance of different class outliers. In the parameter optimization process of this case, the model prefers to increase the general penalty C substantially rather than adjusts the proportion of the cost-sensitive penalty C_{+1}, C_{-1}, if so, according to the formula (1), cost-sensitive penalty C_{+1}, C_{-1} will be ineffective, and the separating hyperplane will be set in the middle of the two classes of training data. The true distribution of the totality, however, may not be like the training dataset. In the case that we pay more attention to the minority class, we need some redundancy to the minority class for robustness. It is said by the terminology of SVM that increasing the margin of the minority class. The ideal separating hyperplane may not be the middle of the training dataset, so the CS-SVM refer to [13] and used in our work can solve this problem.

In making up a clinic prescription, doctors always pay more attention to the positive class, which means the recall of the majority class is more important. They hope to prescribe more rare herbs instead of afraid of making mistakes and refusing to generate them.

In the derivation procedure of standard SVM, we hope the distance between the sample dot closest to separating hyperplane and the separating hyperplane itself, that is to say, the *margin*, will be as large as possible. The distance from a sample x_i to the separating hyperplane can be represented as $w^T x + b$, but the hyperplane is not changed by scaling the w. For simplicity, we use x^* to represent the dot closest to the separating hyperplane and fix $w^T x^* + b = 1$, therefore the margin can be written as $\frac{1}{||w||}$ and we can adjust the margin by scaling w.

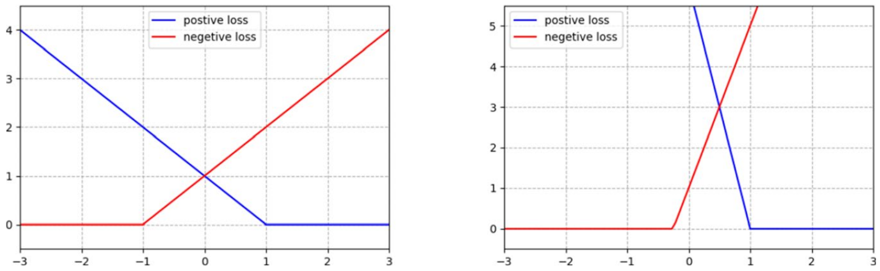
In standard SVM and BP-SVM, we set both the margin of majority class and minority class are the $\frac{1}{||w||}$, but in CS-SVM, we use the κ to replace the numerator 1 to shrink the margin, that is $\frac{\kappa}{||w||}$. We can easily found the free degree in BP-SVM is two that the general penalty C and the proportion between C_{+1}, C_{-1}. However, the CS-SVM has three free degrees, the C_{-1}, on the one hand, can play the role of majority negative class penalty, such as the



(a) Comparison between standrad SVM and BP-SVM

(b) Comparison between BP-SVM and CS-SVM

Fig. 4 Comparison of the separating hyperplane of different SVM



(a) Hinge loss function of standard SVM and BP-SVM

(b) Hinge loss function of CS-SVM

Fig. 5 Comparison of the hinge loss function of different SVM

$\lambda = 2C_{-1} - 1$ in formula (18), on the other hand, it also can shrink the margin of majority class, because of the condition in (15) and the (18), the $\kappa \leq 1$, the margin of majority class will less than the counterpart in minority class forever regardless of whether separable or not. In Figure 4(b), we can easily found that the separating hyperplane of CS-SVM is closer to the negative class and the margin of the positive class is larger.

It can be seen in Figure 5(a) that the turning point of standard SVM loss is (1,0) and (-1,0), so the margin is $\frac{1}{||w||}$. In Figure 5(b), we set $C_{+1} = 6, C_{-1} = 2.5$. Therefore the slope is the 6 and 4, the margin is the $\frac{1}{||w||}$ and $\frac{1}{4*||w||}$ respectively. The slope controls the tolerance of different classes, and the margin ensures the cost-sensitive mechanism can be executed in the situation of data separable by means of forcing shrink the margin of negative class.

4 Experiment

In this section, we conduct extensive experiments to show the effectiveness, efficiency of the proposed method.

4.1 Methods and data sets for performance comparison

To evaluate the performance of the proposed method *BCC+CS-SVM* , we take a combination of multi-label classification algorithm and basic classifier algorithm for comparison.

- Bayes Classifier Chain + Cost Sensitive SVM (BCC+CS-SVM)
- Binary Relevance + Cost Sensitive SVM (BR+CS-SVM)
- Binary Relevance + Biased Penalties SVM (BR+BP-SVM)
- Bayes Classifier Chain + standard SVM (BCC+standard SVM)
- Binary Relevance + standard SVM (BR+standard SVM)
- Seq2seq model based RNN

Table 2 Statistics of the dataset

Quantity of total sample	Input feature	Output labels
10052	189	357

- Herb-Know

where BR+CS-SVM, BR+BP-SVM and BR+standard SVM adopt Binary Relevance (BR) as multi-label classification algorithm, and their basic classifier algorithms are CS-SVM, BP-SVM, and standard SVM, respectively. The purpose of these three experiments is to compare the performance of these three SVMs. BCC+CS-SVM and BCC+standard SVM adopt Bayes Classifier Chain (BCC) as multi-label classification algorithm. These two groups of experiments can be compared with the previous three groups to further illustrate the superiority of the BCC algorithm. In addition, we also compare two deep learning-based methods achieving the task of generating TCM prescriptions, which are Seq2seq model based RNN [9] and Herb-Know [8]. Seq2seq model based RNN uses two different RNN as encoder and decoder. The encoder first takes in a set of symptoms and compresses them into hidden states. With masking and coverage mechanism, the decoder then produces a group of herbs based on the information embodied in the hidden states given by the encoder. Similar to Seq2seq model based RNN, Herb-Know is also a sequence model which adopts Transformer [23] encoder module as its encoder instead of RNN. In decoding, each herb is generated according to both the symptoms and the pre-selected herb candidates.

Our dataset contains more than 10,000 prescriptions of traditional Chinese medicine for lung cancer, all provided by partner hospitals. Our prescription data set D is shown in Table 2. The total number of samples is 10052, where the dimension of the input feature (symptom) is 189, and the dimension of the output label (herbal medicine) is 357. The ratio of the training set to the test set is 9:1.

4.2 Evaluation metrics

In order to test these three support vector machine models on category imbalance data, we classify these labels according to the percentage of positive samples in the total samples. The more the percentage of a label deviates from 50%, the more unbalanced the data in this label. The evaluation indicators we use are commonly used in multi-label classification, such as precision, recall, F1-score, and specificity.

Precision indicates the percentage of correctly predicted positive results among all the predicted positive results. Precision is calculated as follows:

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \tag{19}$$

Recall denotes the proportion of correctly predicted positive results to all actual positive results, which is calculated as follows:

Table 3 The results of initial experiments

	F1-score	Precision	Recall
EasyEnsembleClassifier	0.178	0.237	0.721
BalancedRandomForestClassifier	0.208	0.269	0.702

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \tag{20}$$

F1-score is the harmonic mean of precision and recall taking both metrics into account in the following equation:

$$\text{F1-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{21}$$

Specificity measures the proportion of negatives that are correctly identified, which is calculated using (22).

$$\text{Specificity} = \frac{\text{TrueNegative}}{\text{TrueNegative} + \text{FalsePositive}} \tag{22}$$

Total cost is also used to assess the model's cost sensitivity performance, which is also a zero-one risk of cost sensitivity. In (23) the P_1 and P_{-1} are the class priors probability and P_{FN} and P_{FP} are the false negative and false positive rates respectively.

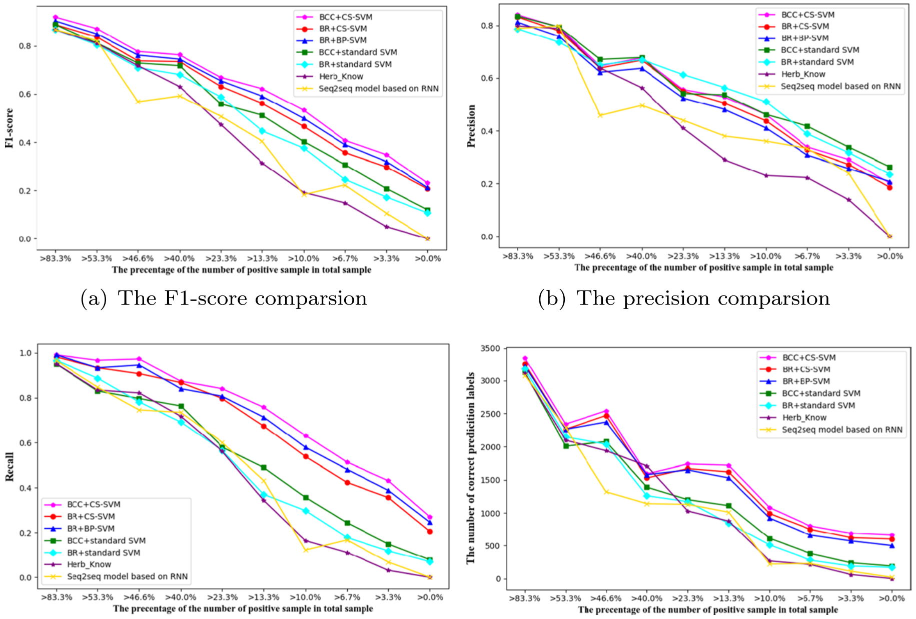
$$\begin{aligned} Totalcost =& E_{\mathbf{Y},\mathbf{X}}[L_{C_{+1},C_{-1}}(f(\mathbf{x}), y)|\mathbf{X} = \mathbf{x}] \\ =& \sum_{y} \sum_{x} P_{\mathbf{X}|\mathbf{Y}}(\mathbf{X} = \mathbf{x}|Y = y) P_{\mathbf{Y}}(y) L_{C_{+1},C_{-1}} f(\mathbf{x}, y) \\ =& \sum_{y} P_{\mathbf{Y}}(+1) \sum_{\mathbf{x}} P_{\mathbf{X}|\mathbf{Y}}(\mathbf{X} = \mathbf{x}|Y = +1) L_{C_{+1},C_{-1}} f(\mathbf{x}, +1) \\ &+ \sum_{y} P_{\mathbf{Y}}(-1) \sum_{\mathbf{x}} P_{\mathbf{X}|\mathbf{Y}}(\mathbf{X} = \mathbf{x}|Y = +1) L_{C_{+1},C_{-1}} f(\mathbf{x}, -1) \\ =& P_1 C_1 P_{FN} + P_{-1} C_{-1} P_{FP} \end{aligned} \tag{23}$$

4.3 Initial experimental analysis

In the initial experiments, we mainly studied sampling related methods to address the problem of class imbalance. In addition to synthetic minority over-sampling technique [3], we mainly used two algorithms based on ensemble learning sampling strategies, namely EasyEnsembleClassifier [11] and BalancedRandomForestClassifier [4]. The main idea of the EasyEnsembleClassifier is to train several classifiers for ensemble learning by repeatedly

Table 4 The evaluation of the total sample

	F1-score	Precision	Recall	Specificity	Total cost
BR+Standard-SVM	0.591	**0.644**	0.545	0.734	19.275
BCC+Standard-SVM	0.604	0.641	0.57	0.748	19.021
BR+BP-SVM	0.606	0.534	0.701	0.821	16.575
BR+CS-SVM	0.618	0.538	0.723	0.834	15.878
BCC+CS-SVM	**0.638**	0.551	**0.757**	**0.853**	**15.743**
Seq2seq model based RNN	0.533	0.551	0.516	0.711	25.763
Herb-Know	0.536	0.541	0.532	0.736	28.175



(a) The F1-score comparsion

(b) The precision comparsion

(c) The recall comparsion

(d) The number of practical correct predicition labels

Fig. 6 Comparison of model performance on different levels of label class imbalance

combining positive samples with the same number of negative samples randomly sampled. The results are shown in the Table 3.

As shown in Table 3, the performance of the sampling strategy method is deficient; that is, there will be exceedingly high false positives, which means that our model will be misled by sampling or synthetic samples in training. As a result, a large number of inappropriate herbs are predicted. A further reason is that the number of samples is not enough. If under a single classification learning scenario, 10,000 samples may be enough. However, for multi-label classification, the number of samples may be too small compared with the predicted dimension, so adopting the methods of changing the sampling strategies and synthetic samples may cause the model to learn wrong information.

4.4 Results in prescription dataset

For all experiments, to avoid random effect and get robust results, we apply 10-fold cross-validation, and the final performance is reported as the average over the ten folds.

Table 4 reports the experimental results of combination of multi-label classification algorithm and basic classifier algorithm. We can find that BCC + CS-SVM has the best performance on the F1-score. Although its precision is lower than the standard BR + standard SVM, it is the tradeoff for expanding the prediction scale to get more correct labels. We also applied two deep learning model Seq2seq model based RNN [9] and Herb-Know [8] on our data. However, the overall evaluation of the deep learning models is lower than that of the SVM models.

Figure 6 shows the comparison between five SVM methods and two deep learning models on classification labels with different imbalances. 357 kinds of labels are classified according to the percentage of positive samples in the total samples. Each portion in Figure 6 do not have a containment relationship, e.g.,> 53.3% for <83.3% and > 53.3% portion.

Combining Table 4 and Figure 6, we get the following observations.

- Regarding the two standard SVM models, in Figure 6, we observe that the BCC algorithm is better than the BR algorithm and has a higher F1-score in Table 4. Also, we find that the BCC algorithm is superior to the BR algorithm in the three cost-sensitive SVM models. These results effectively illustrate that the BCC algorithm can improve the performance of prescription generation models. In other words, when a doctor prescribes, the BCC method can help them consider the relevance of different herbs, such as the classic "The eighteen incompatible medicaments, the nineteen medicaments of mutual restraint".
- In Table 4, we can observe that among the three BR algorithm-based models, although the precision of the model using standard SVM is higher than that of the model using other SVMs, the recall and F1 scores of BP-SVM and CS-SVM are higher. This is a crucial issue. The standard SVM of the evaluation index of the total sample is slightly lower than BP-SVM and CS-SVM, but if we consider the imbalance of each label, in Figure 6(a) (c) (d), we can find that three evaluation matrics on standard SVM, the F1-score, recall and the number of practical correct prediction labels, are significantly lower than the BP-SVM and CS-SVM. If the data of some labels are more unbalanced, the phenomenon is more serious. It is shown that CS-SVM sacrifices part of precision for the improvement of recall, and the overall performance of CS-SVM is better than that of standard SVM under the class imbalance scenario. Specifically, CS-SVM increases the number of predicted positive samples by pushing the separating hyperplane to the negative sample set, which is bound to reduce the precision. However, if more positive samples are predicted through this operation, the recall rate and overall performance will be improved. This is well in accordance with the TCM clinical scenario needs that doctors hope that the model can improve the classification performance of rare herbs as much as possible, rather than ignoring them.
- For the comparison between BP-SVM and CS-SVM, it can be seen from Table 4 that the total cost of CS-SVM is less than BP-SVM. Therefore CS-SVM has better Bayesian consistency, i.e., its minimum conditional risk is closer to the Bayesian error rate. CS-SVM is also higher than BP-SVM in other evaluation metrics, such as F1 scores, recall, and specificity.
- The performance of two deep learning models on certain categories of balanced labels is similar to that of the three SVM models, but as the label imbalance increases, the evaluation is getting lower and lower, and the performance is not as good as the standard SVM. We think this is because the deep learning model relies on large amounts of data, but the number of positive samples with the most unbalanced labels is often less than 100. During training, the small number of samples leads to overfitting of the depth model. However, it is impossible to have a massive number of single-disease prescriptions in actual clinical situations, so deep learning cannot fully use its advantages in this situation.
- For the two deep learning methods, we find that the overall performance of the two methods is similar. However, the performance of Herb-Know is slightly lower than that of the seq2seq model based on RNN when the class imbalance increases. According to our analysis, Herb-Know uses herb pre-selector to pre-select herbs and uses transform-

Table 5 An example for a TCM prescription of lung cancer

Sympotms	放疗中(Radiation therapy)咳嗽 (cough)
	苔白腻(moss white greasy)脉滑 (pulse slip)
	脉数(rapid pulse)神疲(exhausted) 乏力 (lacking in strength)
		Recall

Experts prescription	鱼腥草 (Cordate Houttuynia)蛇舌草 (Hedyotis diffusa) 石见穿 (Salvia chinensis)女贞子 (Ligustrum lucidum Ait) 陈皮(Pericarpium Citri Reticulatae) 石上柏 (Selaginella doederleinii Hieron) 北沙参 (Radix Glehniae)制半夏 (processed Rhizoma Pinelliae) 黄芪 (Astragali radix)苍术 (Rhizoma Atractylodis) 蜀羊泉 (Solanum septemlobum Bunge)僵蚕 (Stiff Silkworm) 麦冬 (Radix Ophiopogonis)浙贝母 (Thunberg Fritillary Bulb) 稻芽 (Rice-grain Sprout)枸杞子 (Fructus Lycii) 猫爪草 (Ranunculus ternatus Thunb) 桑寄生 (Taxillus chinensis)天南星 (Arisaema amurense Maxim) 金沸草 (Herba Inulae)百部 (Radix Stemonae) 蒺藜 (Puncturevine Caltrop Fruit)厚朴 (Magnolia officinalis) 天葵子 (Semiaquilegia adoxoides)	

BR+ Standard SVM	**鱼腥草(Cordate Houttuynia)蛇舌草(Hedyotis diffusa)**
	石见穿(Salvia chinensis)女贞子(Ligustrum lucidum Ait)
	陈皮(Pericarpium Citri Reticulatae)
	石上柏(Selaginella doederleinii Hieron)
	北沙参(Radix Glehniae)制半夏(processed Rhizoma Pinelliae)
	黄芪(Astragali radix)山慈菇 (Cremastra appendiculata)
	枸杞子(Fructus Lycii)杏仁 (Prunus armeniaca)
	菟丝子 (Semen Cuscutae)**麦冬(Radix Ophiopogonis)**
	桔梗 (Platycodon grandiflorum)仙灵脾 (Herba Epimedii)
	金银花 (Lonicera japonica)大枣 (Ziziphus jujuba)
	苍术(Rhizoma Atractylodis)象贝母 (As fritillaria)生米仁 (Coix Seed)
		0.5

BCC+ CS-SVM	**鱼腥草(Cordate Houttuynia)蛇舌草(Hedyotis diffusa)**
	石见穿(Salvia chinensis)女贞子(Ligustrum lucidum Ait)
	陈皮(Pericarpium Citri Reticulatae)
	石上柏(Selaginella doederleinii Hieron)
	北沙参(Radix Glehniae)制半夏(processed Rhizoma Pinelliae)
	黄芪(Astragali radix)山慈菇 (Cremastra appendiculata)
	菟丝子 (Semen Cuscutae)杏仁 (Prunus armeniaca)**僵蚕(Stiff Silkworm)**
	桔梗 (Platycodon grandiflorum)**苍术(Rhizoma Atractylodis)**
	枸杞子(fructus lycii)麦冬(Radix Ophiopogonis)
	稻芽(Rice-grain Sprout)象贝母 (As fritillaria)
	蜀羊泉(Solanum septemlobum Bunge)生米仁 (Coix Seed)
	猫爪草(Ranunculus ternatus Thunb)百部(Radix Stemonae)
		0.708

er's encoder to encode symptoms. Both operations require many samples for training. When the imbalance degree of label increases, its performance will gradually be lower than that of RNN based method.

4.5 Case study

We evaluate our model through an example of a clinical prescription. Table 5 is a specific Chinese medicine prescription that shows the symptom section. The first row of Table 5 is the prescription from the partner hospital's expert. The second and last rows are the prescriptions obtained using the BCC+SVM and BCC+CS-SVM algorithm, respectively. We evaluate the overall performance of prescriptions by recall metric. It can be found that the recall rate of prescriptions obtained by the BCC+CS-SVM algorithm is significantly higher than that obtained by BCC+ Standard SVM. We compare the prescriptions given by these algorithms. In Table 5, we have highlighted in bold the same herbs as those prescribed by the expert. It is noted that the order of the herbs is arranged in the order of the number of times they appear (i.e., the common herbs at the front and the rare herbs at the back). It can be found that several herbs predicted by BCC+CS-SVM are concentrated in the latter part, which also indicates that BCC+CS-SVM has a better prediction capacity for rare herbs than BR+Standard SVM.

We further analyze the prediction errors and find that the medicinal effect of '山慈菇' (Cremastra appendiculata) was similar to that of '天南星' (Arisaema amurense Maxim), both of which have the effect of dispersing nodules and reducing swelling. The medicinal effect of '菟丝子' (Semen Cuscutae) is also similar to that of '蒺藜' (Puncturevine Caltrop Fruit), both of which have the effect of nourishing the liver and brightening the eyes. For'象贝母' (As fritillaria), it is a representative of high quality '浙贝母' (Thunberg Fritillary Bulb). This means that although our proposed model does not exactly match the expert prescription, the model can predict some herbs with similar functions and thus provide doctors with a more flexible choice of herbal combinations (Table 5).

5 Conclusion

TCM has a long and sparkling history and is the most important complement to modern medicine. However, the treatment process of Chinese medicine lacks the standardization of modern medicine. In this paper, We construct a TCM prescription generation model by combining the Bayesian classifier chain algorithm (BCC) and cost-sensitive support vector machine (SVM) to address the relevance and category imbalance problems in the TCM clinical prescriptions. Among them, the BCC method is improved based on the characteristics of traditional Chinese medicine prescriptions, and cost-sensitive modifications such as offset penalty and hinge loss correction are added to the standard support vector machine. These modifications have achieved better performance in our TCM clinical data set. However, there is still room for improvement in this model. For example, the correlation between Chinese medicines is complicated. Maybe we can try other better methods to mine these relationships and adapt our model to more real and complex clinical situations. In addition, some herbs have not been predicted in the case study, and the recall needs to be further improved. Furthermore, the dosage data in the prescription is not fully utilized. In future work, we hope to introduce more TCM knowledge to further improve the performance of TCM prescription generation.

Acknowledgements This work is supported by the National Science Foundation of China (No. 61672161).

References

1. Akbani, R., Kwek, S., Japkowicz, N.: Applying support vector machines to imbalanced datasets. In: European conference on machine learning, pp. 39–50. Springer(2004)
2. Bach, F.R., Heckerman, D., Horvitz, E.: Considering cost asymmetry in learning classifiers. Journal of Machine Learning Research **7**(Aug), 1713–1741 (2006)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
4. Chen, C., Liaw, A., Breiman, L.: Using random forest to learn imbalanced data. Technical Report 666, Department Statistics, UC Berkley (2004). https://doi.org/statistics,berkley.edu/tech-reports/666
5. Dembczynski, K., Cheng, W., Hüllermeier, E.: Bayes optimal multilabel classification via probabilistic classifier chains. In: ICML (2010)
6. Gonçalves, E.C., Plastino, A., Freitas, A.A.: A genetic algorithm for optimizing the label ordering in multi-label classifier chains. In: 2013 IEEE 25th International Conference on Tools with Artificial Intelligence, pp. 469–476. IEEE (2013)
7. Kubat, M., Matwin, S., et al.: Addressing the curse of imbalanced training sets: one-sided selection. In: Icml, vol. 97, pp. 179–186. Nashville, USA (1997)
8. Li, C., Liu, D., Yang, K., Huang, X., Lv, J.: Herb-know: Knowledge enhanced prescription generation for traditional chinese medicine. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1560–1567. IEEE (2020)
9. Li, W., Yang, Z., Sun, X.: Exploration on generating traditional chinese medicine prescription from symptoms with an end-to-end method. arXiv:1801.09030 (2018)
10. Liu, R., Hou, W., Hua, B.J., et al.: Chinese herbal decoction based on syndrome differentiation as maintenance therapy in patients with extensive-stage small-cell lung cancer: an exploratory and small prospective cohort study. Evid.-Based Complement. Alternat. Med. **2015** (2015)
11. Liu, X..Y.., Wu, J.., Zhou, Z..H..: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **39**(2), 539–550 (2008)
12. Ma, J., Wang, Z.: Discovering syndrome regularities in traditional chinese medicine clinical by topic model. In: International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, pp. 157–162. Springer (2016)
13. Masnadi-Shirazi, H., Vasconcelos, N., Iranmehr, A.: Cost-sensitive support vector machines. arXiv:1212.0975 (2012)
14. Pei, C., Ruan, C., Zhang, Y., Yang, Y.: Bayes classifier chain based on svm for traditional chinese medical prescription generation. In: Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data, pp. 748–763. Springer (2020)
15. Read, J., Martino, L., Olmos, P.M., Luengo, D.: Scalable multi-output label prediction: From classifier chains to classifier trellises. Pattern Recognition **48**(6), 2096–2109 (2015)
16. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Machine learning **85**(3), 333 (2011)
17. Ruan, C., Ma, J., Wang, Y., Zhang, Y., Yang, Y.: Discovering regularities from traditional chinese medicine prescriptions via bipartite embedding model. In: International Joint Conferences on Artificial Intelligence, pp. 3346–3352 (2019)
18. Ruan, C., Wang, Y., Zhang, Y., Ma, J., Chen, H., Aickelin, U., Zhu, S., Zhang, T.: Thcluster: herb supplements categorization for precision traditional chinese medicine. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 417–424. IEEE (2017)
19. Ruan, C., Wang, Y., Zhang, Y., Yang, Y.: Exploring regularity in traditional chinese medicine clinical data using heterogeneous weighted networks embedding. In: International Conference on Database Systems for Advanced Applications, pp. 310–313. Springer (2019)
20. Sucar, L.E., Bielza, C., Morales, E.F., Hernandez-Leal, P., Zaragoza, J.H., Larrañaga, P.: Multi-label classification with bayesian network-based chain classifiers. Pattern Recognition Letters **41**, 14–22 (2014)
21. Torre, L., Bray, F., Siegel, R.L., Ferlay, J., Lortet-Tieulent, J.: Global cancer statistics, 2012. CA: A Cancer Journal for Clinicians **65**(2), 87–108 (2015)
22. Tsoumakas, G., Vlahavas, I.: Random k-labelsets: An ensemble method for multilabel classification. In: European conference on machine learning, pp. 406–417. Springer (2007)

23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv:1706.03762 (2017)
24. Wu, G., Chang, E.Y.: Adaptive feature-space conformal transformation for imbalanced-data learning. In: Proceedings of the 20th International Conference on Machine Learning (ICML-03), pp. 816–823 (2003)
25. Xu, Q., Tang, W., Teng, F., Peng, W., Zhang, Y., Li, W., Wen, C., Guo, J.: Intelligent syndrome differentiation of traditional chinese medicine by ann: A case study of chronic obstructive pulmonary disease. IEEE Access **7**, 76167–76175 (2019)
26. Yao, L., Zhang, Y., Wei, B., Zhang, W., Jin, Z.: A topic modeling approach for traditional chinese medicine prescriptions. IEEE Transactions on Knowledge and Data Engineering **30**(6), 1007–1021 (2018)
27. Zhang, T., et al.: Statistical behavior and consistency of classification methods based on convex risk minimization. The Annals of Statistics **32**(1), 56–85 (2004)