Bayes Classifier Chain Based on SVM for Traditional Chinese Medical Prescription Generation

Chaohan Pei[1,4], Chunyang Ruan[2], Yanchun Zhang[3,4(✉)], and Yun Yang[5]

[1] School of Software Engineering, Fudan University, Shanghai, China
17212010027@fudan.edu.cn

[2] School of Economics and Finance Shanghai International Studies University, Shanghai, China
cyruan16@fudan.edu.cn

[3] College of Engineering and Science, Victoria University, Melbourne, Australia

[4] Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, China
Yanchun.Zhang@vu.edu.au

[5] Department of Oncology and Longhua Hospital, Shanghai, China

Abstract. Traditional Chinese Medicine (TCM) plays an important role in the comprehensive treatment of lung cancer. However the quality of the prescriptions from TCM doctors depends on the doctor's personal experience, which leads to the TCM prescriptions are the lack of standardization. We apply the original clinical TCM prescriptions data to train a standardized prescription generating model for TCM therapy. Our model adopts the Bayes Classifier Chain (BCC) algorithm to solve the label correlation problem, whose basic classifier is cost-sensitive SVM targeted to the class imbalance of the label. The results of experiments on the prescription dataset demonstrated the effectiveness and practicability of the proposed model for a prescription generation.

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

1 Introduction

As one of the most common malignant tumors, lung cancer is a leading cause of cancer-related death worldwide [15]. TCM is considered as an important complementary therapy with beneficial effects for lung cancer patients by reducing toxic effects, improving the quality of life [8]. It can be observed that traditional Chinese medicine has become an important part of the comprehensive treatment system for lung cancer. However, different from the normalized diagnosis and treatment standard in modern medicine, traditional Chinese medicine is more

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

X. Wang et al. (Eds.): APWeb-WAIM 2020, LNCS 12317, pp. 748–763, 2020.
https://doi.org/10.1007/978-3-030-60259-8_55

individualized in the treatment of patients, and the treatment effect is closely linked to the doctor's level of clinical experience. For example, the prescription made by TCM doctors, which consists of a set of herbs, may be different from different doctors. Therefore, it is a very meaningful task to integrate the clinical prescriptions of different TCM doctors, analyze the rules, and then standardize the prescribing process. In relevant research about the TCM standardization, the prescription data were mostly from the TCM classics and pharmacopeia. There is just an obvious problem with these datasets that the prescriptions of TCM medical books are too old and simple to suit the up-to-date medical demand. Fortunately, our collaboration hospital has provided over 10000 prescriptions of TCM therapy aimed at lung cancer and we applied these data in our experiment.

Table 1 shows an example of TCM prescription excerpted from an electronic medical record. The first row is the set of symptom descriptions. The practitioner prescribes herbs shown in the second row based on the symptoms and diagnosis. In this paper, we construct a multi-label classifier, whose input is a set of symptoms and the output is a group of herbs.

Table 1. An example for a TCM 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)

In our early experiment, we found two critical problems with the prescription dataset. The first problem is the correlation between each herb label, for example, each prescription has a fundamental prescription targeted to a specific symptom comprised of several fixed herbs. These herbs often appear together in a certain prescription. The second is the class imbalance of every label interior. In our total dataset, there are 357 herb labels, 189 symptom features, and 10000+ samples. However, there are 255 labels in total whose the number of positive samples only accounts for less than 3.3% in total samples. It can be seen in Fig. 1 that the number of positive samples with most of the labels is much less than the negative samples.

In multi-label classification algorithms, the Binary Relevance (BR) [3] is the basic algorithm, which converts the multi-label classification to single-label classification to solve. BR algorithm is simple and doesn't consider the label correlation, but the reality is complicated. Aimed to the label correlation, the label power-set approach [16] transforms the multidimensional problem into a single-class scenario by defining a new compound class variable whose possible values are all of the possible combinations of values of the original classes. It is the obvious disadvantage of this method that the computational complexity will increases exponentially with the number of labels. Based on BR, Classifier Chains (CC) algorithm [12] constructs a chain structure on labels and determines the presence/absence of the current label under the condition of previously determined

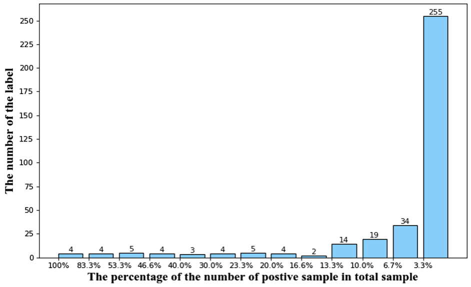


Fig. 1. We have 357 labels in our prescription dataset and we classify this label according to the percentage of the number of positive samples in the total sample. The height of every bar is the actual quantity of each category

labels. There are some problems with CC methods such as how to decide the order of the labels in the chain, and not all labels exist the correlation between each other.

Zaragoza [14] proposed a more effective method, Bayes Classifier Chain (BCC), which establishes a directed acyclic graph of the label set based on the correlation between each label. Then they train each classifier starting from the top node, the results of the parent node classifier will be added into the input feature set of the children node classifier. In our work, we built a DAG for the herb label set according to the special attribute of the prescription dataset refer to the BCC method to solve the labels correlation problem.

The solution to cope with the class imbalance can roughly be grouped into two general categories. The first is to address the problem from the respect of sampling, that is to say changing the distribution of the sample, by adopting resampling techniques such as oversampling, undersampling and synthetic sampling with data generation [1,4,6]. In our previous experiment, the performance after altering the sampling strategy was dissatisfactory because of the abnormally high false positive. Therefore we adopt the second category solution in this paper. This method is called cost-sensitive learning using different cost matrices that describe the costs for misclassifying any particular data example. In our research, we selected the SVM as the basic classifier of CC method and modified the SVM by cost-sensitive means.

We refer to the work of Masnadi-Shirazi et al. [10], in which they proposed a new cost-sensitive SVM. This new model not only can deal with the class imbalance problem but also implemented the cost-sensitive Bayes decision rule and made the model risk approximate the cost-sensitive Bayes risk. The experiment result showed that the performance of this SVM is better than others. In the following sections, we call this SVM as CS-SVM (cost-sensitive SVM). The contributions of our work are as follows:

- We improve the BCC method targeted to the unique feature about the TCM prescriptions dataset. In our BCC classifier, the DAG construction approach exhibits the fine interpretability of TCM prescriptions.

- We combine the multi-label learning algorithm with the cost-sensitive SVM and compare its performance with other different SVM algorithm. This CS-SVM exhibits excellent performance in dealing with a class imbalance of label interior in multi-label classification problems.
- We apply our multi-label classification model on the TCM prescription prediction problem and achieve better performance, which was approved by TCM doctors.

2 Related Work

2.1 TCM Knowledge Discovery

With the development of artificial intelligence, more researches pay attention to the TCM data mining using AI. The topic model has been widely applied in the analysis of the prescriptions, such as Jialin Ma et al. [9], Liang Yao et al. [19]. The graph theory model also plays an important role in TCM research. Chunyang Ruan et al. [13] adopted the graph model to find the rule between symptoms and herbs in TCM. With the development of deep learning, more and more researchers tried to adapt the neural network method into biomedical to deal with medical problems. Wei Li et al. [7] proposed a seq2seq model based on RNN to generate the herbs, which refer to the machine translation model in NLP. Qiang Xu et al. [18] chose chronic obstructive pulmonary disease as an example of investigating syndrome differentiation for TCM based on artificial neural networks.

2.2 Classifier Chain

Read et al. [12] first introduced chain classifiers as an alternative method for multi-label classification that incorporates class dependencies, while keeping the computational efficiency of the binary relevance approach. Based on the fundamental CC method, researchers have done many improvements. Dembczynski proposed [5] Probabilistic Chain Classifier (PCC) algorithm, which mainly applies a probability frame in CC. Although PCC can better consider the relativity between labels, it has very high time complexity. Goncalves et al. [14] referred to the genetic algorithm and then put forward the GACC algorithm, the purpose is to optimize the CC forecast order chain by the heuristic algorithm. J. Read et al. [11] presented the classifier trellis (CT) method for scalable multi-label classification. In recent work, we can see that many researchers pay close attention to the label order by searching for the correlation between the labels.

2.3 Cost Sensitive SVM

SVMs are based on a very solid learning-theoretic foundation and have been successfully applied to many classification problems. The cost-sensitive modification on the basic SVM algorithm can cope with the class imbalance problem and there two primary cost-sensitive modifications on SVM. The first was known as the biased penalties SVM (BP-SVM) [2,17], whose mechanism consists of applying different penalty factors C_1 and C_{-1} for the positive and negative SVM slack variables during training. It is implemented by transforming the primal SVM problem into

$$\begin{aligned} &arg \min_{w,b,\xi} \frac{1}{2}||w||^2 + C\left[C_1 \sum_{\{i|y_i=1\}} \xi_i + C_{-1} \sum_{\{i|y_i=-1\}} \xi_i\right] \\ &\text{s.t. } y_i(w^T x + b) \geq 1 - \xi_i \end{aligned} \tag{1}$$

The BP-SVM suffers from an obvious flaw, which has limited ability to carry out a cost-sensitive strategy when the training data are separable. In the process of parametric optimization, the model intends to select large slack penalty C rather than adjust the cost-sensitive penalty C_1 and C_{-1} and then the slack variable ξ is zero-valued and the optimization degenerates into that of the standard SVM, where the separating hyperplane is placed midway between the two classes (rather than assigning a larger margin to one of them). The second is a cost-sensitive SVM model proposed by [10] and in this paper, we call it CS-SVM for simplicity. They modified the hinge loss function directly by the cost-sensitive way rather than only added penalty terms. We will elaborate on it in the following section.

3 Methodology

Our prescription predicting can be regarded as a multi-label classification mission. In the following, we use the boldface to represent a vector and the normal font is the scalar or a component of a vector. Every train sample consists of a symptom set and a herb set, which can be represented as $(\mathbf{X}_i, \mathbf{Y}_i)$. $\mathbf{X}_i$ is the input vector and $\mathbf{Y}_i$ is the output vector, in our problem, they are deemed as symptom vector and herb vector respectively. For every $\mathbf{X}_i = [x_1, x_2, \cdots, x_M] \in \{-1, 1\}^M$,$\mathbf{Y}_i = [y_1, y_2, \cdots, y_L] \in \{-1, 1\}^L$, the M and L are the dimension of the input vector and output vector severally. In our task, M is the number of total symptoms and the L is the number of total herbs. If the symptom set of one sample contains a symptom s_j, $(j = 1, \cdots, M)$, the jth component of the vector, x_j, will be 1, otherwise will be -1 and the herb set is like the symptom set. Our task is training a multi-label classifier $F(\cdot)$ satisfied the functional relationship $\mathbf{Y} = F(\mathbf{X})$ on the basis of training sample.

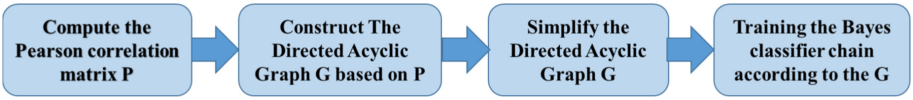

Compute the Pearson correlation matrix P
Construct The Directed Acyclic Graph G based on P
Simplify the Directed Acyclic Graph G
Training the Bayes classifier chain according to the G

Fig. 2. The framework of our BCC training procedure

3.1 Bayes Classifier Chain Algorithm

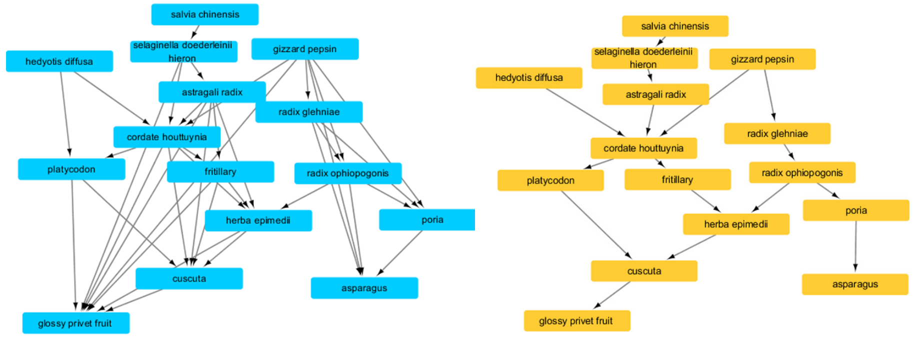
Algorithm 1 and Fig. 2 are the framework of our BCC algorithm. The BCC in this research has two parts:

1. Constructing the order of the classifier chain, the directed acyclic graph,
2. Training the BCC classifier according to the DAG and this part is elaborated in Algorithm 3.

Construct the Directed Acyclic Graph(DAG). Ordinarily, constructing a Bayes network is an NP-hard problem. However, we can simplify this process based on the dataset feature in our prescriptions generation task.

Firstly, we count the occurrence number of every herb label in total 10000+ sample, and then sort all the labels by their occurrence numbers from large to small. We find that if a herb's occurrence frequency is higher, it will be more important and common use when doctors make a prescription. TCM doctors always consider the common herbs at first and then judge whether to use rare herbs. This fact means that we can set the direction of the herb network from the high-frequency herbs to low-frequency herbs and the most common herbs are start nodes in this network. In label sample matrix H, where $H \in \{-1, +1\}^{N \times L}$, the column vector $\mathbf{y_i}$ are arranged by the herb frequency order above.

Secondly, we compute the Pearson correlation coefficient matrix $P(\mathrm{L} \times \mathrm{L})$ between every herbs label based on the label sample matrix H. Every element $p_{i,j} \in [0, 1]$ in P, after consultation with the doctor, we select the correlation coefficient threshold: 0.2 after many experiments, if $|p_{i,j}| > 0.2$, we regard the herb y_i and y_j exist correlation, and then $p_{i,j} = 1$ ortherwise $p_{i,j} = 0$ (Fig. 3).



(a) Original DAG of herb labels (b) DAG after network pruning by DFS

Fig. 3. The node-set of these two networks consists of the top 15 highest occurrence frequency herb labels in the total sample. The previous node in the topological sorting order of the network has a higher frequency than the later node.

Thirdly, we construct a DAG $G =< V, E >$, where V is the node set and every node v_i corresponding to a herb label y_i. We stipulate that if $p_{i,j} = 1$ and $i > j$, then the directed edge$< y_i, y_j >\in E$. At last, this DAG exists a problem

Algorithm 1. Framework of BCC training for our system.

```
Input: The sample matrix of symptom features $S$, where $S \in \{-1, +1\}^{N \times M}$. The
    sample matrix of herb labels $H$, where $H \in \{-1, +1\}^{N \times L}$.;
Output: The BCC classifier $F$;
1: Compute the Pearson correlation coefficient matrix of each label $P$,where $P \in$
   $\{-1, +1\}^{L \times L}$ according to $S$;
2: Select 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 get the
   adjacent matrix $G$ based on $P$.
3: Apply the Algorithm 2, input the DAG adjacent matrix $G$, get simplified DAG
   adjacent matrix $G'$.
4: Use the DAG: $G' =< V, E' >$ as the classifier chain order of BCC algorithm, then
   call Algorithm 3.
```

that if there is a path $r_{i,j}$ from v_i to v_j, we can find there are many directed edge $< v_k, v_j >$, where $\{v_k | v_k \in r_{i,j}\}$. Targeted to this problem, we apply deep first search(DFS) algorithm to remove redundant edge, which is explained in Algorithm 2.

Bayes Classifier Chain. If the DAG has been established, the training process is following Algorithm 3. We chose some special options for general training procedures. Firstly, in the training process, if we want to train a classifier for label y_i, we select the actual class value of the ancestor label node about the y_i given in the original training set instead of the prediction value in training, which will tend to produce more accurate classifiers. Secondly, we use all ancestor nodes of the label that will be training as the additional input features besides the symptoms, because this scheme conforms to the general way of thinking for TCM doctors.

Algorithm 3 shows the training procedure in detail. This algorithm references the DFS algorithm and makes some modification. We ensure that if a label node will be training, all of its ancestor nodes have been ended their training process. When we train along one path and counter a node that has more than one in-degree, we will add the additional feature of this node's parent node in the path and decrease this node's in-degree by one. Then we start from other paths until this node's in-degree equal zero. If so, we can continue from this node.

3.2 Cost Sensitve SVM

The BCC can solve the label correlation problem in our prescription generation task to some degree. But there still exists the class imbalance problem, so we will improve our model in the aspect of the basic classifier. We selected the cost-sensitive SVM [10] as the basic classifier for BCC algorithm.

Bayes Consistent of Standard Binary Classifier. For binary-classify task, the goal is to predict an ubobserve value $y \in \{+1, -1\}$ based on an observed input vector $\mathbf{x}$. This requires us to train a functional relationship $y = h(\mathbf{x})$ from a set of example pairs of $(\mathbf{x}, y)$. From a statistical viewpoint, the feature vectors

Algorithm 2. Simplify the DAG of herb labels node.

```
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 $D'$
```

and class labels can be regarded as random variable possessing probability distributions $P_X(\mathbf{x})$ and $P_Y(y)$ respectively. We write the classifier function as the form that $h(\mathbf{x}) = sign[p(\mathbf{x})]$, where the function $p : \mathcal{X} \to \mathbb{R}$. A non-negative function $L(p(\mathbf{x}), y)$ be deemed as the loss function for each $(p(\mathbf{x}), y)$ pair. The classifier is considered optimal if it minimizes the expected loss $R = E_{\mathbf{X},Y}[L(p(\mathbf{x}), y)]$, also known as the expected risk. Minimizing the expected loss also equivalent to minimizing the conditional risk

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

To make it easier to understand this formula in probability way, we can write the predictor function $p(\mathbf{x})$ as a composition of two functions $p(\mathbf{x}) = f(\eta(\mathbf{x}))$, where $\eta(\mathbf{x}) = P_{Y|\mathbf{X}}(1|\mathbf{x})$ is the posterior probability. $f : [0, 1] \to \mathbb{R}$ is called the link function in this paper, which establishes a connection to Bayes decision rule by this means. The Bayes error rate of the data distribution is the probability that an instance is misclassified by a classifier which knows the true class probabilities given the predictors. We hope minimized *conditional risk* closed to the Bayes error. Assuming the true probability distribution has been known, if we want to minimize the conditional risk, we can select the suited link function f when the loss function L is fixed.

The ϕ is the concrete form of the loss function L, such as the hinge loss in SVM $\phi(yf) = \lfloor 1 - yf \rfloor_+$,where $\lfloor x \rfloor = max(0, x)$. The f is the function of η, but for simplicity, we omit the η. Because the loss function ϕ may be different in a false positive and false negative, these cost-sensitive loss function can also be written as a unified form

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 $F(\cdot) = \left[f_0, f_1, f_2, \ldots, f_{num(v)-1}\right]$
```

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

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

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

There exists a suitable link function $f^*_\phi(\eta)$ and it can minimized the conditional risk $C_{\phi,C_1,C_{-1}}$.

Cost Sensitive SVM Loss Function. In this section, we will expand the hinge loss function to cost-sensitive version. The loss function of standard SVM is hinge loss, $\phi(yf) = \lfloor 1 - yf \rfloor_+$,where $\lfloor x \rfloor = max(0, x)$. Refer to [20] , the optimal link function for standard SVM is

$$f^*_\phi(\eta) = sign(2\eta - 1) \tag{5}$$

and the minimum conditional risk is

$$\begin{aligned} &C^*_\phi(\eta) = 1 - -2|2\eta - 1| \\ &= \eta\lfloor 1 - sign(2\eta - 1)\rfloor_+ + (1 - \eta)\lfloor 1 + sign(2\eta - 1)\rfloor_+ \end{aligned} \tag{6}$$

We modify the optimal link function of standard SVM by cost-sensitive parameter naturally

$$f^*_{\phi,C_1,C_{-1}}(\eta) = sign((C_1 + C_{-1})\eta - C_{-1}) \tag{7}$$

Like the conditional risk of standard SVM $C^*_\phi(\eta)$, we get the cost-sensitive counterpart

$$\begin{aligned} C^*_{\phi,C_1,C_{-1}}(\eta) =& \eta\lfloor e - d \cdot sign((C_1 + C_{-1})\eta - C_{-1})\rfloor_+ + \\ & (1 - \eta)\lfloor b + a \cdot sign((C_1 + C_{-1})\eta - C_{-1})\rfloor_+ \end{aligned} \tag{8}$$

where

$$d \geq e, \quad a \geq b, \quad \frac{C_{-1}}{C_1} = \frac{a + b}{d + e} \tag{9}$$

and the a, b, d, e are positive number. Then we can easily find that

$$sign((C_{-1} + C_1)\eta - C_{-1}) = \begin{cases} 1, & \text{if } \eta \geq \gamma \\ 0, & \text{if } \eta = \gamma \\ -1, & \text{if } \eta \leq \gamma \end{cases} \tag{10}$$

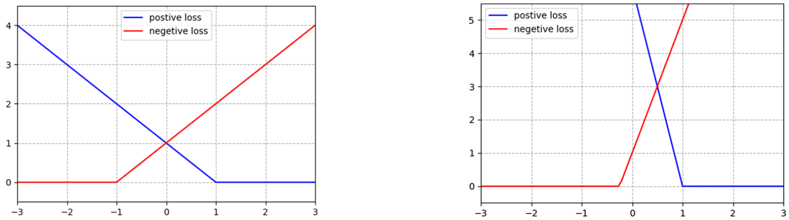
where $\gamma = \frac{C_{-1}}{C_1 + C_{-1}}$. If $\eta < \gamma$, the risk is

$$C^*_{\phi,C_1,C_{-1}}(\eta) = \eta\lfloor e + d\rfloor_+ + (1 - \eta)\lfloor b - a\rfloor_+ \tag{11}$$

At last, like the form of the hinge loss about standard SVM, the loss function of cost-sensitive SVM can be deduced

$$\phi_{C_1,C_{-1}}(yf) = \begin{cases} \lfloor e - df\rfloor_+, & \text{if } y = 1; \\ \lfloor b + af\rfloor_+, & \text{if } y = -1; \end{cases} \tag{12}$$

There are four freedom degrees in this hinge loss function, which control the margin and slope of two class respectively. The positive class divide by margin $\frac{e}{d}$ and slope d of hinge loss and the negative class divide by margin $\frac{b}{a}$ and slope a of hinge loss.



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

(b) Hinge loss function of CS-SVM

Fig. 4. The (a) is the hinge loss of standard SVM and BP-SVM, $\phi(yf) = max(0, 1 - yf)$. The (b) is the hinge loss of CS-SVM, where $C_{+1} = 6, C_{-1} = 2.5, \lambda = 2C_{-1} - 1 = 4$,positive loss is $\phi_{+1}(yf) = max(0, 6-6yf)$, negative loss is $\phi_{-1}(yf) = max(0, 1+4yf)$.

Cost Sensitive SVM Algorithm. In the previous section, the loss function of cost-sensitive SVM have four freedom degree but in fact, there are only two freedom degree in conditional risk function $C^*_{\phi,C_1,C_{-1}}$. We can find that we only need the proportional relation between the two class slope, $\frac{e}{d}$ and $\frac{b}{a}$. So we suppose that the positive class weight is more important, which requires the slope and margin of positive class is higher than the counterpart of negative class,

$$\frac{e}{d} \geq \frac{b}{a} \quad d \geq a \tag{13}$$

and then fix the $\frac{e}{d} = 1$ and set $e = d = C_1$ in order to specify the postive class margin. In a similar way, we only need the proportional relation between the a and b. The b can be set at 1, the accord to the third folumation of (9), $a = 2C_{-1} - 1$. At last, we bring the value of a, b, d, e into (8) and obtain the resulting cost sensitive SVM minimal conditional risk is

$$\begin{aligned} C^*_{\phi,C_1,C_{-1}}(\eta) =& \eta \lfloor C_1 - C_1 \cdot sign((C_1 + C_{-1})\eta - C_{-1}) \rfloor_+ + \\ & (1-\eta) \lfloor 1 + (2C_{-1} - 1) \cdot sign((C_1 + C_{-1})\eta - C_{-1}) \rfloor_+ \end{aligned} \tag{14}$$

with $C_{-1} \geq 1$,$C_1 \geq 2C_{-1} - 1$ in order to satisfy (13). The intuitional explanation is that the positive class has a larger margin that can make the separating hyperplane deviated to negative class and have a higher slope can increase the cost risk when occurring misclassification.

There the standard SVM risk can be modified by the cost-sensitive method:

$$\begin{aligned} & arg \min_{w,b} \sum_{\{i|y_i=1\}} \lfloor C_1 - C_1(w^T x_i + b) \rfloor_+ \\ & + \sum_{\{i|y_i=-1\}} \lfloor 1 + (2C_{-1} - 1)(w^T x_i + b) \rfloor_+ + \mu ||w||^2 \end{aligned} \tag{15}$$

then deduce to a primer optimization problem

$$\begin{aligned} & arg \min_{w,b,\xi} \frac{1}{2}||w||^2 + C\left[\beta \sum_{\{i|y_i=1\}} \xi_i + \lambda \sum_{\{i|y_i=-1\}} \xi_i\right] \\ & \text{s.t. } \; y_i(w^T x + b) \geq 1 - \xi_i, y_i = 1 \\ & \qquad y_i(w^T x + b) \geq \kappa - \xi_i, y_i = -1 \end{aligned} \tag{16}$$

with

$$\beta = C_1 \quad \lambda = 2C_{-1} - 1 \quad \kappa = \frac{1}{2C_{-1} - 1} \tag{17}$$

In this quadratic programming problem, the cost-sensitivity is controlled by the three parameters β, γ, κ. The β, γ decide the relative weights of margin violations and pay more attention to positive class on the constraint that $C_{-1} \geq 1$,$C_1 \geq 2C_{-1} - 1$. When the data are separated, the BP-SVM(1) has a defect that the optimization procedure tends to select larger the parameter C in BM-SVM(1), in that circumstances, the cost-sensitive parameter C_1, C_{-1} will be ineffective and degenerate into standard SVM. But in this model, the κ can shrink to narrow the margin rather than increase the common slack penalty C. The Fig. 4 shows that the distinction of loss function between standard SVM, BP-SVM, and CS-SVM. The SVM's margin is the X-intercept and the X-intercept of BP-SVM is 1 and −1, but the negative loss X-intercept's absolute value of CS-SVM is smaller than 1.

4 Experiment

In this section, we conduct several experiments to compare the performance of CS-SVM with BP-SVM and standard SVM with BR and BCC algorithm. We implement our method by the SVM library libsvm[1].

Table 2. The prescription dataset

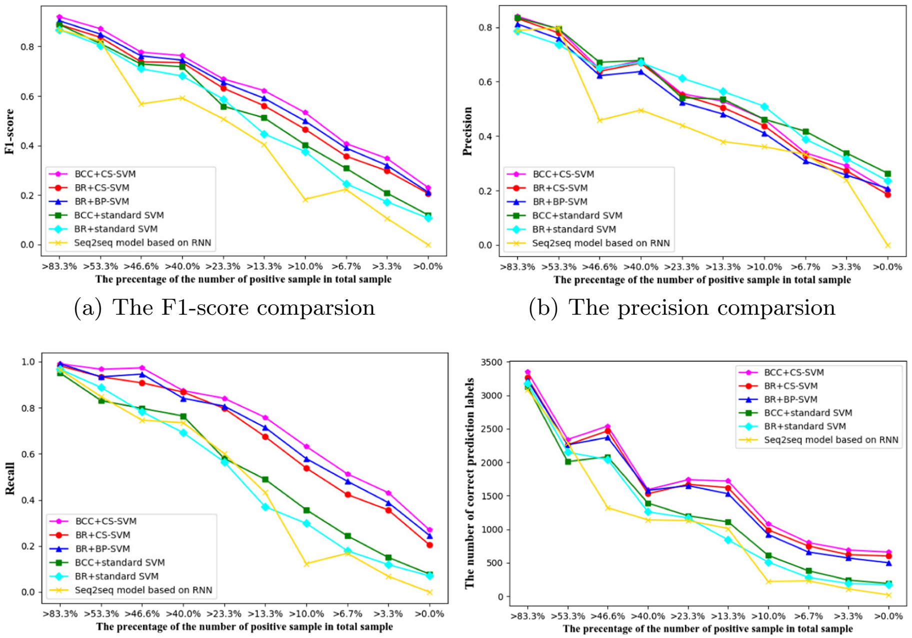
Quantity of total sample	Input feature	Output labels
10052	189	357

4.1 Dataset

Our dataset consists of 10000+ TCM prescriptions targeted to lung cancer, which were provided by the cooperative hospital. Our prescription dataset D has been shown in Table 2. The quantity of the total sample is 10052, in which the dimension of input feature(the symptom) is 189 and the number of the output labels(the herbs) is 357. The proportion of training set to test set is 9:1. The data will be upload our github[2].

[1] https://www.csie.ntu.edu.tw/cjlin/libsvm/.

[2] https://github.com/xbybshd/TCM-prescription-dataset.



(a) The F1-score comparsion

(b) The precision comparsion

(c) The recall comparsion

(d) The number of practical correct predicition labels

Fig. 5. The multi-label evaluation of the samples which are classified by the degree of imbalance of every label, we applied BR+standard SVM, BCC+standard SVM, BR+BP−SVM, BR+CS−SVM, BCC+CS−SVM and a Seq2seq RNN model.

4.2 Multi-label Classifiy Evaluation Index

To test these three SVM models on class imbalance data, we classify these labels according to the percentage of the number of positive samples in the total sample. The percentage of a label is more deviate 50%, the data in this label are more imbalance. The evaluation indexes we adopt are the common metrics in multi-label classification, such as precision, recall, specificity, F1-score, and G-means. To test the performance in the practical application more clearly, we statistics the raw number of the label in prediction set, validation set and the intersection of prediction and validation. Besides, the total cost is also be applied to evaluate the cost sensitivity of the model, which is also the cost-sensitive zero-one risk. The $Totalcost = P_1 C_1 P_{FN} + P_{-1} C_{-1} P_{FP}$, where 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.

4.3 Result in Prescription Dataset

Table 3 is the global evaluation of three SVM methods on our test dataset and we can find that the BCC+CS−SVM exhibits the best on F1-score. Although the precision is lower than standard BR+standard SVM, it is the tradeoff for expanding the prediction scale to obtain more correct labels. We also applied a recent deep-learning model proposed by Wei Li [7] on our data, which is based on the RNN seq2seq model for a prescription generation. But the total evaluation of the RNN seq2seq model is lower than the SVM models.

Figure 5 is the comparison between five SVM methods and the Seq2seq RNN model on the labels classified by different degrees of imbalance. The 357 labels are classified by the percentage of themselves a positive sample in the total sample. Each part in the Fig. 5 doesn't have the inclusion relation, for example, that the ¿53.3% represents the part of ¡83.3% and ¿53.3%.

Table 3. The evaluation of the total sample

	F1-score	Precision	Recall	G-means	Total cost
BR+Standard-SVM	0.591	**0.644**	0.545	0.734	null
BCC+Standard-SVM	0.604	0.641	0.57	0.748	null
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	null

In two standard SVM model, we can find the model used BCC algorithm is slightly better than the BR algorithm in Fig. 5, and the BCC model(green line) also have higher total F1-score in Table 3. The Similar situation also appears in three cost-sensitive SVM model. These phenomena prove that our BCC algorithm can improve the performance of the prescription generation model. When the doctor makes a prescription, the BCC method can consider the correlation between the herbs compared with the BR method, such as the classical rule that "The eighteen incompatible medicaments, the nineteen medicaments of mutual restraint".

In the three BR algorithm model, although the precision of BR+standard SVM is higher than others in Table 3, the BP-SVM and CS-SVM have higher recall and F1-score. It is a critical problem that the evaluation index on the total sample of the standard SVM is slightly lower than BP-SVM and CS-SVM, but if we consider the imbalance degree of every labels, in Fig. 5(a)(c)(d), we can find the three evaluation on standard SVM, the F1-score, recall and the number of practical correct prediction labels, are obviously less than counterparts of BP-SVM and CS-SVM. This phenomenon is more serious if the data of certain labels are more imbalance, which confirms the previous analysis that the standard SVM will push the separating hyperplane to the minority class and results in a few predictions. In practical application, doctors hope the model to pay attention to the minority label rather than omit them simply.

As for the comparison between BP-SVM and CS-SVM, the fifth column of Table 3 is the comparison of the total cost between the BP-SVM and CS-SVM and the cost of CS-SVM is less than BP-SVM. The total cost is also the expectable risk of the model, the lower risk indicates the error rate of this model is more close to the Bayes error rate. In other evaluation indexes, such as F1-score, the CS-SVM is also higher than the BP-SVM.

The performance of the RNN Seq2seq model is similar to the three SVM models in some class balance labels but the evaluation becomes lower and lower with the increase of the label unbalancedness, which performs worse than the standard SVM. We believe this result is because the deep learning model relies on mass data, but the number of the most unbalanced label positive sample often less than 100. In the training process, the small quantity of the sample makes the deep model overfit. But it is impossible that there is an enormous quantity of the single-disease prescriptions in the practical clinical situation so the deep learning can't exert its advantage in this situation.

5 Conclusion

TCM is one of the most significant complementary and alternative medicine and it plays an important role in the therapy of lung cancer. However, the therapeutic process of TCM lacks standardization like modern medical. Targeted to the herb correlation and class imbalance problem in clinical TCM prescription, we combined the Bayes classifier chain algorithm (BCC) and cost-sensitive SVM to process the firsthand clinical TCM prescription and construct a simple TCM prescription generating model. In detail, the BCC method was modified based on the feature of TCM prescriptions and the cost-sensitive modifications were added on standard SVM, such as bias penalty and amending the hinge loss. These modifications have obtained better performance in our clinical TCM dataset. But this model still has some room for improvement, for example, the correlations between the herbs are complex, and maybe we can try other better methods to mine these relationships to make our model adapt the more real complex clinical situation.

References

1. Akbani, R., Kwek, S., Japkowicz, N.: Applying support vector machines to imbalanced datasets. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML 2004. LNCS (LNAI), vol. 3201, pp. 39–50. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30115-8_7
2. Bach, F.R., Heckerman, D., Horvitz, E.: Considering cost asymmetry in learning classifiers. J. Mach. Learn. Res. **7**(Aug), 1713–1741 (2006)
3. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recogn. **37**(9), 1757–1771 (2004)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
5. Cheng, W., Hüllermeier, E., Dembczynski, K.J.: Bayes optimal multilabel classification via probabilistic classifier chains. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 279–286 (2010)

6. Kubat, M., et al.: Addressing the curse of imbalanced training sets: one-sided selection. In: Icml, vol. 97, pp. 179–186. Nashville, USA (1997)
7. Li, W., Yang, Z., Sun, X.: Exploration on generating traditional Chinese medicine prescription from symptoms with an end-to-end method. arXiv preprint (2018). arXiv:1801.09030
8. Liu, R., 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. Altern. Med. **2015** (2015)
9. Ma, J., Wang, Z.: Discovering syndrome regularities in traditional Chinese medicine clinical by topic model. 3PGCIC 2016. LNDECT, vol. 1, pp. 157–162. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-49109-7_15
10. Masnadi-Shirazi, H., Vasconcelos, N., Iranmehr, A.: Cost-sensitive support vector machines. arXiv preprint (2012). arXiv:1212.0975
11. Read, J., Martino, L., Olmos, P.M., Luengo, D.: Scalable multi-output label prediction: from classifier chains to classifier trellises. Pattern Recogn. **48**(6), 2096–2109 (2015)
12. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Mach. Learn. **85**(3), 333 (2011)
13. Ruan, C., Wang, Y., Zhang, Y., Yang, Y.: Exploring regularity in traditional Chinese medicine clinical data using heterogeneous weighted networks embedding. In: Li, G., Yang, J., Gama, J., Natwichai, J., Tong, Y. (eds.) DASFAA 2019. LNCS, vol. 11448, pp. 310–313. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-18590-9_35
14. 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 Recogn. Lett. **41**, 14–22 (2014)
15. Torre, L.A., Bray, F., Siegel, R.L., Ferlay, J., Lortet-Tieulent, J., Jemal, A.: Global cancer statistics, 2012. CA Cancer J. Clin. **65**(2), 87–108 (2015)
16. Tsoumakas, G., Vlahavas, I.: Random *k*-labelsets: an ensemble method for multilabel classification. In: Kok, J.N., Koronacki, J., Mantaras, R.L., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 406–417. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74958-5_38
17. 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)
18. 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)
19. Yao, L., Zhang, Y., Wei, B., Zhang, W., Jin, Z.: A topic modeling approach for traditional chinese medicine prescriptions. IEEE Trans. Knowl. Data Eng. **30**(6), 1007–1021 (2018)
20. Zhang, T., et al.: Statistical behavior and consistency of classification methods based on convex risk minimization. Ann. Stat. **32**(1), 56–85 (2004)