Analysis of TCM Diagnosis and Treatment of Thyroid Diseases Based on Data Mining

Kankan She and Caiyan Dai*

School of artificial intelligence and information technology, Nanjing University of Traditional Chinese Medicine, Nanjing, China

*Corresponding author e-mail: daicaiyan@njucm.edu.cn

Abstract. Thyroid disease is the most common and frequently occurring endocrine disease. In recent years, the incidence rate of thyroid diseases has increased in recent years due to the gradual change of the ecological environment and the gradual popularization of iodine supplements for iodine deficiency disorders. At present, the prevalence of hyperthyroidism is 1.3%, hypothyroidism is 6.5%, thyroid nodules are 18.6%, and 2%-5% of thyroid nodules are thyroid cancer. Thyroid disease has become a high incidence and common disease in modern society. The number of thyroid disease patients in China has become the second largest disease after hypertension. Thyroid disease belongs to gall disease of traditional Chinese medicine. Compared with western medicine, traditional Chinese medicine has certain advantages in the treatment of gall disease. Compared with western medicine, traditional Chinese medicine can better improve clinical symptoms and reduce toxicity.

In this paper, Stata and python were used to analyze and clean the data of TCM diagnosis thyroid medical records. Principal component analysis and factor analysis were used to analyze and clean the symptoms and pathogenesis of the medical record data. The data in the prescription were quantified. The feature selection of the data set was performed by principal component analysis to reduce the data dimension. In the process of treatment, different kinds of prescriptions will be used. Association analysis algorithm in data mining is used to mine the association rules of prescription medication and complications of thyroid disease, which provides a reference for clinical treatment drug selection and disease prevention of traditional Chinese medicine. Logistic regression was used to establish the prediction model of traditional Chinese medicine diagnosis prescription of thyroid by using the clinical data of traditional Chinese medicine.

Keywords: Thyroid Disease, TCM Diagnosis, Traditional Chinese Medicine, Prescription Analysis, Logistic Regression, Principal Component Analysis

1. Introduction

Thyroid diseases are common diseases in Department of Endocrinology. The incidence rate is increasing year by year, mainly including hyperthyroidism, hypothyroidism, thyroiditis, thyroid
nODULES, THYROID TUMORS, ETC. ACCORDING TO THE EPIDEMIOLOGICAL DATA OF THYROID DISEASES AMONG URBAN RESIDENTS IN CHINA, THE PREVALENCE OF HYPOTHYROIDISM AND THYROID NODULES IS 17.8% AND 18.6% RESPECTIVELY. HYPERTHYROIDISM ACCOUNTED FOR 1.6% [1-3], AND THERE WERE MORE THAN 200 MILLION PATIENTS WITH THYROID DYSFUNCTION IN CHINA, AND THE PREVALENCE RATE SHOWED A RAPID UPWARD TREND. MODERN MEDICINE MAINLY USES ANTI-THYROID DRUGS, HORMONE REPLACEMENT, RADIATION THERAPY OR SURGERY IN THE TREATMENT OF THYROID DISEASES. ANTI-THYROID DRUGS OR HORMONES CAN QUICKLY CONTROL THE THYROID FUNCTION INDEX IN THE NORMAL RANGE, BUT IT IS EASY TO RELAPSE AFTER WITHDRAWAL OF DRUGS, AND THE SIDE EFFECTS ARE LARGE, SO THE CLINICAL SYMPTOMS OF THYROID DISEASE ARE NOT IMPROVED SIGNIFICANTLY. ACCORDING TO TRADITIONAL CHINESE MEDICINE, THYROID DISEASE BELONGS TO THE CATEGORY OF "GALL DISEASE". THE SCOPE OF "GALL DISEASE" IS VERY WIDE. IT IS CAUSED BY THE ACCUMULATION OF QI STAGNATION, PHLEGM OBSTRUCTION AND BLOOD STASIS IN THE ANTERIOR LARYNGEAL NODE OF THE NECK. THE TYPICAL MANIFESTATION IS GALL SWELLING OR MASS FORMATION IN FRONT OF THE NECK [4-6]. IN RECENT YEARS, CHINESE SCHOLARS HAVE REPORTED A LARGE NUMBER OF RESEARCH RESULTS ON THE TREATMENT OF THYROID DISEASES WITH TRADITIONAL CHINESE MEDICINE [5], WHICH CONFIRMED THAT TRADITIONAL CHINESE MEDICINE HAS ADVANTAGES IN REGULATING HUMAN ENDOCRINE, REGULATING BODY IMMUNITY, ELIMINATING ANTIBODIES AND STABILIZING PITUITARY GLAND. AMONG THEM, THE CURATIVE EFFECT OF FAMOUS TRADITIONAL CHINESE MEDICINE IN THE DIAGNOSIS AND TREATMENT OF THYROID DISEASES IS PARTICULARLY SIGNIFICANT. IN THIS PAPER, TCM DIAGNOSIS AND TREATMENT OF THYROID MEDICAL RECORD DATA CLEANING AND MINING, USING PRINCIPAL COMPONENT ANALYSIS, FACTOR ANALYSIS OF MEDICAL RECORD DATA IN THE SYMPTOMS, PATHOGENESIS ANALYSIS, ESTABLISH DISEASE PREDICTION MODEL, GUIDE CLINICAL MEDICATION.

2. DATA AND METHODS

2.1. RESEARCH OBJECT

The data of 998 medical records collected from Guoyitang of Nanjing University of traditional Chinese medicine include age, gender, diagnosis of traditional Chinese medicine, classification of Western medicine, diagnosis of Western medicine, clinical manifestations, tongue coating, pulse condition, physical and chemical examination, pathogenesis, rule of law, prescription, remarks and diagnosis time.

2.2. RESEARCH METHODS

The medical records were imported into pychar, and the frequency of drug use and frequency of drug use in 998 patients who met the inclusion criteria were counted with Python 3.7, and then the data were imported into stata16 for further analysis. According to the frequency of use, the top 15 herbs were selected. Principal component analysis, correlation analysis and factor analysis were used to analyze and clean the symptoms and pathogenesis in the medical record data. The data in the prescription were quantified. The feature selection of the data set was performed by principal component analysis to reduce the data dimension [6-8].

3. RESEARCH RESULTS

3.1. SINGLE DRUG ANALYSIS

Rhizoma Pinelliae and Prunella vulgaris are the most commonly used drugs in clinical practice, which are 748 times, accounting for about 75% of the total cases. Prunella vulgaris can clear liver fire, level liver Yang, soothe liver depression and disperse phlegm and nuclear. Combined use of the two drugs, a drop of a powder, cold and warm, clearing heat and phlegm, treatment of liver qi stagnation, long and fire, phlegm and heat mutual knot of insomnia, wipe phlegm core, thin tumor. The drugs with more frequent occurrence are shown in Table 1.
Table 1. High frequency medication

| NO. | Drug                           | Frequency | Proportion  | NO. | Drug                           | Frequency | Proportion  |
|-----|--------------------------------|-----------|-------------|-----|--------------------------------|-----------|-------------|
| 1   | Pinellia ternata               | 803       | 74.67%      | 17  | Fried Eucommia ulmoides        | 255       | 23.73%      |
| 2   | Prunella                       | 803       | 74.67%      | 18  | Pollen Typhae                  | 232       | 21.60%      |
| 3   | Rhizoma cyperi                 | 683       | 63.47%      | 19  | Peony bark                     | 227       | 21.07%      |
| 4   | Tangerine peel                 | 608       | 56.53%      | 20  | Fried xuanhu                   | 227       | 21.07%      |
| 5   | Fired Sinapis Alba             | 499       | 46.40%      | 21  | Curcuma kwangtungensis         | 227       | 21.07%      |
| 6   | Radix Paoniae Alba             | 459       | 42.67%      | 22  | Fritillaria thunbergii         | 227       | 21.07%      |
| 7   | Scorched Atractylodes          | 453       | 42.13%      | 23  | White Poria cocos              | 218       | 20.27%      |
| 8   | Roasted licorice               | 424       | 39.47%      | 24  | Arrowhead mushroom             | 209       | 19.47%      |
| 9   | Ophiopogon japonicus          | 381       | 35.47%      | 25  | Gouju plum                     | 204       | 18.93%      |
| 10  | Vinegar Bupleurum chinense    | 370       | 34.40%      | 26  | Sedum sarmentosum              | 204       | 18.93%      |
| 11  | Fried Fructus aurantii        | 364       | 33.87%      | 27  | Caulis Spatholobus             | 201       | 18.67%      |
| 12  | Radix Scrophulariae            | 353       | 32.80%      | 28  | Angelica sinensis              | 201       | 18.67%      |
| 13  | Ligusticum chuanxiong         | 344       | 32.00%      | 29  | Radix Pseudostellariae         | 198       | 18.40%      |
| 14  | Codonopsis pilosula           | 324       | 30.13%      | 30  | Stir fried Atractylodes lancea | 192       | 17.87%      |
| 15  | Loranthus yadoriki Sieb       | 287       | 26.67%      | 31  | Calcined longmuge              | 186       | 17.33%      |
| 16  | Radix salviae miltiorrhizae   | 270       | 25.07%      | 32  | Gardenia heishanensis         | 181       | 16.80%      |

3.2. Drug Association Analysis

Combined with machine learning method, APRIORI association algorithm and data cleaning technology in Python, the original data in traditional Chinese medicine prescriptions were standardized, and the drug compatibility relationship was mined by association rule model. The compatibility relationship is as follows:

\[
\begin{align*}
&\text{([{'Fired Sinapis Alba'}, {'Prunella'}, {'Tangerine peel'}, {'Pinellia ternata'}, {'Rhizoma cyperi'}])}, \\
&\text{([{'Prunella'}, {'Pinellia ternata'}]), ([{'Rhizoma cyperi'}, {'Prunella'}, {'Pinellia ternata'}]), ([{'Rhizoma cyperi'}, {'Pinellia ternata'}])}], \\
&\text{([{'Tangerine peel'}])} \Rightarrow ([{'Prunella'}, {'Pinellia ternata'}]) \text{ conf: 0.8571428571428571} \\
&\text{([{'Pinellia ternata'}])} \Rightarrow ([{'Rhizoma cyperi'}, {'Prunella'}]) \text{ conf: 0.7285714285714285} \\
&\text{([{'Prunella'}])} \Rightarrow ([{'Rhizoma cyperi'}, {'Pinellia ternata'}]) \text{ conf: 0.8773584905660378} \\
&\text{([{'Tangerine peel'}])} \Rightarrow ([{'Prunella'}, {'Pinellia ternata'}]) \text{ conf: 0.9760765550239233} \\
&\text{([{'Rhizoma cyperi'}, {'Prunella'}])} \Rightarrow ([{'Pinellia ternata'}]) \text{ conf: 0.9389678414096916} \\
&\text{([{'Prunella'}, {'Pinellia ternata'}])} \Rightarrow ([{'Rhizoma cyperi'}]) \text{ conf: 0.8192771084337349} \\
&\text{([{'Pinellia ternata'}, {'Tangerine peel'}])} \Rightarrow ([{'Prunella'}]) \text{ conf: 0.9029126213592233} \\
&\text{([{'Prunella'}, {'Pinellia ternata'}])} \Rightarrow ([{'Tangerine peel'}]) \text{ conf: 0.7469879518072289} \\
&\text{([{'Prunella'}, {'Tangerine peel'}])} \Rightarrow ([{'Pinellia ternata'}]) \text{ conf: 0.9789473684210526} \\
\end{align*}
\]
3.3. **Principal Component Analysis**

There is a certain correlation between the variables. When there is a certain correlation between the two variables, it can be explained that the two variables reflect the information of the subject has a certain overlap. Principal component analysis is to delete redundant repetitive variables (closely related variables) and establish as few new variables as possible, so that these new variables are unrelated, and these new variables keep the original information as far as possible in reflecting the information of the subject.

Principal component analysis (PCA) \cite{9-11}, or principal component analysis (PCA), is a statistical method that tries to combine the original variables into a group of new independent comprehensive variables, at the same time, according to the actual needs; a few less comprehensive variables can be taken out to reflect the information of the original variables as much as possible. In this modeling process, many kinds of symptoms were found, so the dimension of 164 kinds of symptoms was reduced by principal component analysis, so as to better analyze in the future.

Where eigenvalue is the eigenvalue, proportion is the contribution rate, and cumulative is the cumulative contribution rate. The results are shown in Table 2. The prediction results are shown in Figure 1.

### Table 2. Principal component factor analysis

| Component | Eigenvalue | Difference | Proportion | Cumulative  |
|-----------|------------|------------|------------|-------------|
| Comp1     | 7.901      | 1.752      | 0.050      | 0.050       |
| Comp2     | 6.149      | 0.405      | 0.039      | 0.089       |
| Comp3     | 5.745      | 0.108      | 0.037      | 0.126       |
| Comp4     | 5.637      | 0.277      | 0.036      | 0.162       |
| Comp5     | 5.360      | 0.253      | 0.034      | 0.196       |
| Comp6     | 5.107      | 0.244      | 0.033      | 0.229       |
| Comp7     | 4.863      | 0.497      | 0.031      | 0.260       |
| Comp8     | 4.366      | 0.087      | 0.028      | 0.287       |
| Comp9     | 4.280      | 0.142      | 0.027      | 0.315       |
| Comp10    | 4.138      | 0.039      | 0.026      | 0.341       |
| Comp11    | 4.099      | 0.011      | 0.026      | 0.367       |
| Comp12    | 4.088      | 0.046      | 0.026      | 0.393       |
| Comp13    | 4.042      | 0.119      | 0.026      | 0.419       |
| Comp14    | 3.923      | 0.455      | 0.025      | 0.444       |
| Comp15    | 3.468      | 0.025      | 0.022      | 0.466       |

**Figure 1.** Factor analysis

3.4. **Logistic Regression Analysis**

Logistic regression is a generalized linear model \cite{12} so it has many similarities with multiple linear
regression analysis. Their model forms are basically the same, with \( w^x + b \), in which \( w \) and \( B \) are parameters to be determined, and the difference lies in their different dependent variables. Multiple linear regression directly takes \( w^x + b \) as dependent variable, i.e., \( y = w^x + b \), while logistic regression uses function \( L \) to correspond \( w^x + b \) to a hidden state \( p \), \( p = L(w^x + b) \), and then determines the value of dependent variable according to the size of \( P \) and \( 1-p \). If \( L \) is a logistic function, it is a logistic regression; if \( L \) is a polynomial function, it is a polynomial regression. Logistic regression was performed for the 16 most frequently used drugs\(^{[13]}\). The results are shown in Table 3.

### Table 3. Logistic regression

| var37 | Coef. | St.Err. | t-value | p-value | [95% Conf Interval] | Sig |
|-------|-------|---------|---------|---------|---------------------|-----|
| f3    | -0.429| 0.124   | -3.45   | 0.001   | -0.672 -0.185       | *** |
| f6    | 0.389 | 0.133   | 2.92    | 0.004   | 0.128 0.649         | *** |
| f21   | 0.813 | 0.238   | 3.42    | 0.001   | 0.347 1.278         | *** |
| f44   | -0.632| 0.216   | -2.93   | 0.003   | -1.056 -0.209       | *** |
| Constant | 2.035 | 0.206   | 9.88    | 0.000   | 1.631 2.438         | *** |

Mean dependent var 0.816 SD dependent var 0.388
Pseudo r-squared 0.246 Number of obs 343.000
Chi-square 80.644 Prob > chi2 0.000
Akaike crit. (AIC) 256.522 Bayesian crit. (BIC) 275.711

*** \( p<0.01 \), ** \( p<0.05 \), * \( p<0.1 \)

According to the standard that the probability of drug use is at least 0.5, 275 times are accurate, that is, the drug was used. In the other 19 prescriptions, the prediction probability of the model was less than 0.5, and there was no medication. Therefore, the total classification accuracy is 275+19=294 divided by 343, which is 85.71%. The table also provides some conditional probabilities, such as sensitivity, or, in the case of medication, the percentage of cases with a prediction probability greater than or equal to 0.5 (275 out of 280, or 98.21%).

According to table 3, the prediction model is \( P=1[1+\exp(2.035+0.429f3-0.389f6-0.813f21+0.632f44)] \), and the model is 1 with Prunella vulgaris, otherwise it is 0. The prediction results are shown in Figure 2.

![Figure 2. Prunella vulgaris prediction](chart.png)

4. Result Discussion

Prunella vulgaris is bitter and pungent in taste and cold in nature. Its function is to clear the liver, relieve fire and disperse swelling. It is one of the important medicines for the treatment of gall disease in ancient times. "Ben Jing" refers to "dispersing gall and forming Qi", while "materia medica Congxin" states that it "governs scrofula, rat fistula and gall tumor". Pinellia is pungent and slightly warm in nature, and its function is to remove dampness and dissipate phlegm. "BIE Lu" says that
"eliminate carbuncle and swelling", and "drug record" says "can remove gall". Clinical good use of Pinellia ternata, take its longer than dry and wet and lower temperature, avoid consumption of Qi and Yin. In the treatment of hyperthyroidism, Prunella vulgaris is good at clearing away heat and purging fire, and Pinellia ternata is good at resolving phlegm and dispersing nodules. The two have the functions of clearing heat, resolving phlegm and dispersing nodules, which coincides with the pathogenesis of hyperthyroidism based on Yin deficiency and hyperactivity of fire, qi stagnation, phlegm and blood stasis, so it has good clinical effect. In addition, the cold and cool of Prunella vulgaris can help to control the temperature and dryness of Pinellia ternata to avoid damaging Qi and Yin; the pungent powder of Pinellia vulgaris can also avoid the stagnation of Qi and blood caused by cold and cool of Prunella vulgaris.

The model of this study mainly aims at the prediction of whether the drugs used frequently in the treatment of thyroid diseases are used, and gives some answers to the question of whether the corresponding drugs are related, which can provide certain reference and assistance for doctors in the process of intelligent diagnosis and treatment of thyroid diseases.

Acknowledgments
This work was financially supported by National Natural Science Foundation of China (Youth program, No.82004498), Natural Science Foundation of MJUCM( NZY82004498), Jiangsu Province Science Foundation for Youths (No.BK20180822)

References
[1] Le Zhou, Gianlorenzo Dionigi, Alessandro Pontin, Antonella Pino, Ettore Caruso, et al. How does neural monitoring help during thyroid sugery for Graves’ disease [J]. Journal of Clinical & Translational Endocrinology, 2019, 15.
[2] Qikun Zhang, Xiaomin Wang, Junling Yuan, Lu Liu, Ruifang Wang, Hong Huang, Yuanzhang Li. A hierarchical group key agreement protocol using orientable attributes for cloud computing [J]. Information Sciences, 2019, 12(3):480-483
[3] Fred A., De Marsico M., Advances in pattern recognition applications and methods [J]. Neurocomputing, 2018, 173(1): 1-2
[4] Kiran R. U., Kitsuregawa M., Reddy P. K., Efficient discovery of periodic-frequent patterns in very large databases [J]. Journal of Systems and Software, 2018, 112(1): 110-121
[5] Borkar V., Carey M., Grover R., et al, Hyracks: a flexible and extensible foundation for data-intensive computing [C], Proceedings of the 27th IEEE International Conference on Data Engineering (ICDE), 2018: 1151-1162
[6] Nobuyuki Kawahara, Shota Hashimoto, Eiji Tomita. Spark discharge ignition process in a spark-ignition engine using a time series of spectra measurements [J]. Proceedings of the Combustion Institute, 2018, 34: 234-239
[7] Yinsheng Fan, Jicheng Bai, Qiang Li, Research on Maintaining Voltage of Spark Discharge in EDM [J]. Procedia CIRP, 2018, 42: 123-127
[8] Fred A., De Marsico M., Advances in pattern recognition applications and methods [J]. Neurocomputing, 2018, 173(1): 1-2
[9] Kiran R. U., Kitsuregawa M., Reddy P. K., Efficient discovery of periodic-frequent patterns in very large databases [J]. Journal of Systems and Software, 2018, 112(1): 110-121
[10] Wen Sheng Du, Bao Qing Hu. Dominance-based rough set approach to incomplete ordered information systems [J]. Information Sciences, 2018, 346(C):106-129
[11] Shu-Hua Teng, Min Lu, A-Feng Yang, Jun Zhang, Yongjian Nian, Mi He. Efficient attribute reduction from the viewpoint of discernibility [J]. Information Sciences, 2018, 326: 189-197
[12] Wen Chang, Bin Zhu. On rooted cluster morphisms and cluster structures in 2-Calabi–Yau triangulated categories [J]. Journal of Algebra, 2018, 458: 345-350
[13] Jing Zhang, Lan-yue Zhang, Ming Du, Identifying the major air pollutants base on factor and cluster analysis, a case study in 74 Chinese cities [J]. Atmospheric Environment, 2018,
144:37-46