Data Preprocessing and Classification for Taproot Site Data Sets of PANAX NOTOGINSENG

Dao Huang, Jin He*

School of Electronics and Information Engineering Yunnan University of Nationalities, Kunming, China
e-mail: 654874546@qq.com, *Corresponding author: e-mail: hjmwh@sina.com

Abstract—**The herbs from different producing regions have differences in the active constituents and efficacy. The quality of the herb from the authentic region is better than other producing regions. Nowadays, many peddlers substitute non-authentic herbs for authentic-region herbs in order to make more money. So it is important to distinguish herbs between different producing regions. This paper studies the data preprocessing and classification of taproot site data sets of Panax notoginseng from three different producing regions. Compare the effect of data preprocessing includes data standardization, instance selection, attribute selection and try to find out the best method and parameter settings for the data sets. Finally, we use different classification algorithms to classify the preprocessed data and compare the classification performance to find the optimal classification algorithm for the data sets. The classification performance in the experiment was evaluated by Percent Correct (PC), Mean Squared Error (MSE), Kappa Statistics (KS), Area Under ROC (AUR), Mean Absolute Error (MAE). The results shows that using decimal scaling to standardize the data and choose the subset of attribute {1,2,4,6,7,8}is suitable for the data and Random Forest algorithm and AdaBoost.M1 algorithm are the optimal classification algorithm for this data sets which has better classification performance.**

Keywords-AdaBoost.M1, authentic-region herbs, Random Forest; data preprocessing

I. INTRODUCTION

There are different identification methods in the field of herb identification, such as traditional method which identify the herb through their the appearance [1-2], or identify herbs through analyzing the ingredients of herbs that was extracted by the advanced chemical instruments [3-6]. Mining the data extracted by the chemical instruments can make the identification of herb be more effective.

The data preprocessing is one of the most important part in data mining technology. The current related studies in this area have been mainly focused on two aspects: one is the data cleaning and the other is the data reduction. In terms of data cleaning many researchers had studied the anomaly detection [7], the way to clean duplicated records [8-9], and the way to clean the data [10-11]. In terms of data reduction they had studied the reduction of the dimension of high-dimensional data [12], and the discrete technology [13], etc.

At present, research related to herbs' data classification algorithm had focused on the traditional single classification algorithm [14-15], but most single classification algorithm are not suitable to every kinds of data, in order to salve the problem, many scholars had turned to study the multi-classifier algorithm, they had proposed a lot of good multi-classifier algorithms, such as Bagging [16], Boosting [17], AdaBoost[18], Random-Forest[19] ,etc. These multi-classifiers algorithm had been used in many fields and obtained good effect. Therefore, we can use multi-classifier algorithms in the field of herbs' data classification to improve the classification performance. In order to distinguish between different kinds of herbs with different producing regions, this paper firstly preprocess the data extracted by the taproot site of Panax notoginseng, and then find out the optimal data preprocessing method by comparing several evaluation parameters, and then compare several different classification algorithms, choose the better one. This study demonstrates that it is suitable to use decimal scaling standardized data conversion and choose the subset of attribute {1, 2, 4, 6, 7, 8}, the suitable classification algorithms are Random-Forest and AdaBoost.M1.It can be helpful in the study of herbs' data classification.

II. METHOD

A. Data standardization

There are three common standardized operations: Zero-mean standardization, standardization of the decimal scaling, and the Minimum - maximum standardization.

Minimum-maximum standardization (Min-Max):a linear transformation of the data. Suppose S is an attribute and the $\min_S$ is the minimum values of S and the $\max_S$ is the maximum values of S .The Minimum-maximum standardization can be defined as:

$$w=\frac{v-\min_S}{\max_S-\min_S}(a-b)+b \qquad (1)$$

While it mapped v to w , where w is in the interval $[b,a]$, where v is the value of S .

Minimum-maximum standardization will maintain the relationship between the original data values.

Zero-mean standardization (Z-m): it can be defined as:

$$w = \frac{v - \mu_S}{\sigma_S} \qquad (2)$$

Where v is the value of S, w is the standardization result, and μ_S is the mean of S, σ_S is the standard deviation of S.

Decimal Scaling standardization (DS): by moving the decimal place of S. Move the decimal point depends on the maximum absolute value of S . It can be defined as:

$$w = \frac{v}{10^j} \qquad (3)$$

Where j is the smallest integer with the precondition of $Max(|w|) < 1$.

III. Results and discussion

A. Data

The data in this experiment was obtained by Panax notoginseng Institute with the technology of HPLC (High Performance Liquid Chromatography).After analyzing the fingerprint of the data. We can obtain 100 samples of each producing region, and we can totally obtain 300 samples with 3 different regions. One of fingerprints is shown in Fig. 1.

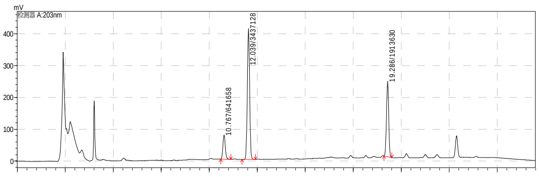

Figure 1. One of fingerprints of taproot site of Panax notoginseng.

B. Comparison of data standardization

In this study, we do the experiment in WEKA. We firstly convert the raw data into the suitable file format such as CSV format and ARFF format. Then we add a class attribute to the file and set 1, 2, 3 as three different regions.

Then we use three kinds of standardization that introduced in 1.1 to process the data. Then compare the three standardization by using several classification algorithms. They are Naive Bayes (BY), Random Forest (RF), Bagging (BA) , NNge (NNG) and RBFnetwork (RN).We set 10-fold cross-validation, the other is the default setting. Then we use Percent Correct (PC), Kappa Statistics (KS), Area Under ROC (AUR),Mean Absolute Error (MAE) as evaluation standard to evaluate the classification performance in order to find the better standardization. The results are shown in TABLE 1, TABLE 2, TABLE 3and TABLE 4.

The results in Tables show that the decimal scaling standardization has better classification performance in comparing the four evaluation standard. So we can choose it for the data standardization.

TABLE I. Comparison of PC after three standardization

	BY	RF	BA	NNG	RN
Z-m	85.27	97.27	92.53	93.9	94.23
Min-Max	85.47	97.13	92.57	93.83	94.43
DS	85.53	97.5	92.57	93.8	94.47

TABLE II. Comparison of KS after three standardization

	BY	RF	BA	NNG	RN
Z-m	0.78	0.96	0.89	0.91	0.91
Min-Max	0.78	0.96	0.89	0.91	0.92
DS	0.78	0.96	0.89	0.91	0.92

TABLE III. Comparison of AUR after three standardization

	BY	RF	BA	NNG	RN
Z-m	0.95	1	0.98	0.94	0.99
Min-Max	0.95	1	0.98	0.94	0.99
DS	0.95	1	0.98	0.94	0.99

TABLE IV. Comparison of MAE after three standardization

	BY	RF	BA	NNG	RN
Z-m	0.11	0.06	0.09	0.04	0.05
Min-Max	0.11	0.06	0.09	0.04	0.05
DS	0.11	0.06	0.09	0.04	0.05

C. Comparison of different attribute subsets

We choose AttributeSelectedClassifier(ASC) to do the attribute selection, ASC is one of classifier in WEKA. We select three different evaluations, they are InfoGainAttributeEval (IGAE), CfsSubsetEval (CSE) and WrapperSubsetEval (WSE).

Firstly, we select Random Forest as the based classifier in ASC, choose IGE as the evaluation, Ranker as the search method. We set numToSlect of Ranker from 7 to 1. Then we number the attribute from 1 to 8 .Every number represent an attribute, from 1 to 8 ,they are R1, Rg1, Rb1, Amount of sample, Moisture content, R1ct, Rg1ct and Rb1ct.The class attribute does not need to be selected .So we can get several subsets of attribute in the first step, they are {7,5,8,6,2,3,1},{7,5,8,6,2,3}, {7,5,8,6,2}, {7,5,8,6}, {7,5,8}, {7,5}, {7}. Secondly, we set the CSE as the evaluation and BestFirst as the search methods. So we can get a subset of attributes {3, 4, 5, 6, 7, 8}. Thirdly, we choose WSE as the evaluation and BestFirst as the search methods. We can get a subset of attributes {1, 2, 4, 6, 7, 8}. We use RF algorithm to classify the different subsets of attribute and then compare their classification performance with four evaluation standard, they are PC,KS,MAE, MSE(Mean square Error).So we can get the results shown in TABLE 5.

As the results shown in TABLE 5, we can draw the conclusion that the attribute subset {1,2,4,6,7,8} can get better classification performance. Therefore, we can use the subset when do the attribute selection.

D. *Comparison of different classification algorithms*

In this experiment, we use naive Bayes(BA), J48, Random Forest (RF), AdaBoost.M1 (ADM), Neural Network (BP), the Nearest Neighbor(k-NN), Support Vector Machine (SVM) and the Vote algorithm which integrate all the seven algorithms introduced before. We set J48 as the based classifier of AdaBoost.M1. We use PC, KS, MAE, MSE, AUR to evaluate the classification performance. We set 10-fold cross-validation, the other is the default setting. Then we can get the results shown in TABLE 6.

TABLE V. Comparison of different subsets of attribute

	{7,5,8,6,2,3,1}	{7,5,8,6,2,3}	{7,5,8,6,2}	{7,5,8,6,}	{7,5,8}	{7，5}	{7}	{3,4,5,6,7,8}	{1,2,4,6,7,8}
PC	98.3539	97.9424	97.9424	97.5309	97.1193	92.5926	67.078	98.3539	99.17
KS	0.9753	0.9691	0.9691	0.963	0.9568	0.8889	0.5062	0.9753	0.9877
MAE	0.0466	0.046	0.0468	0.0494	0.0599	0.0916	0.2518	0.0535	0.0472
MSE	0.104	0.1126	0.1226	0.1324	0.1537	0.219	0.3901	0.1269	0.0948

As the results shown in TABLE 6, we can draw the conclusion that the Random Forest and Adaboost.M1 have the better classification performance, and the Random Forest algorithm do better than Adaboost.M1 in percent correct rate.

TABLE VI. Comparison of different classification algorithms

	BA	J48	RF	ADM	BP	K-NN	SVM	Vote
PC	71.32	94.64	99.76	99.01	90.92	97.44	66.75	97.53
MAE	0.26	0.04	0.05	0.01	0.08	0.02	0.3	0.18
MSE	0.37	0.15	0.09	0.03	0.2	0.09	0.39	0.22
KS	0.57	0.92	1	0.99	0.86	0.96	0.5	0.96
AUR	0.89	0.98	1	1	0.98	0.98	0.93	1

IV. Conclusion

In this paper, we discuss the pre-process and the classification of the data of the taproot site of Panax notoginseng. We firstly study the standardization of the data and find out the suitable standardization method by comparison of several evaluation standard, and the results shown that decimal scaling standardization is suitable to the data in this paper. Then we compare the different attribute subsets selected by three kinds of methods .And we finally find out the suitable subsets by compare several evaluation standard. The subset is {1, 2, 4, 6, 7, 8}.At last ,we study the different classification performance when the data is classified by different classification algorithms, both single classification algorithms and multi-classifier algorithms. We draw the conclusion that Random Forest algorithm and AdaBoost.M1 algorithm are suitable for the data sets .The study in this paper can make a guidance for the pre-processing and classification step of the herbs' data sets identification.

Acknowledgment

Supported by the National Natural Science Foundation of China (61365007/F030406)

References

[1] Yang Jinhua. Gastrodin and identification of its adulterants. [J]. Asia-Pacific Traditional Medicine, 2014, 22: 42.

[2] Jiang shiying.Experience of identification about adulteration of Poria cocos herbs [J] Hubei College of Traditional, 2014, 05: 50-51.

[3] Tu Rong Ling, Wang Jianwei.Research of identification of the authenticity of peel medicine by Near-infrared spectroscopy [J] Chinese folk medicine, 2014, 24: 15-16.

[4] Han Ying, Bifu Jun, Hou Huichan, Zhangyong Yao. Research of application in identification of the authenticity Polygonum multiflorum by near-infrared spectroscopy [J] Chinese Materia Medica, 2014,22: 4394-4398.

[5] He Guangwei new, Li Zhongqiong.Comparison and identification of salviae miltiorrhizae and purple salviae miltiorrhizae [J]. Chinese herbal medicine, 1996,07: 342-343.

[6] Lvxu Yang, Zhang Jizhong, Zhang Zhifeng, Liu Yang,Zengre, Lu Jianmei, Renhuan Ming. Establishment and Identification of Radix Vladimiriae's UPLC characteristic fingerprint [J] Chinese Materia Medica, 2014,14: 2699-2703.

[7] Jeremy Kubica,Andrew Moore,Probabilistic Noise Identification and Data cleaning, Proceedings of the 3rd IEEE International Conference on Data Mining,PP.131-138,2003.

[8] Helena Galhardas,Daniela Florescu,Dennis Shasha,An Extensible Framework for Data cleaning,Proceedings of International Conference on Data Engineering, PP.312, 2000.

[9] Hamid Haidarian Shahri,Ahmad Abdollah Zadeh Barforush,Data Mining for Removing Fuzzy Duplicates Using Fuzzy Inference,Annual Conference Of the North American Fuzzy Information Processing Society,vol.1,PP.419 一 424,2004.

[10] Gediga,G.and Duntsch,I.,Maximum Consistency of Incomplete Data viaNon-Invasive Imputation, in Artificial Intelligence Review, vol.19, no.1, PP.93-107,2003.

[11] Scheffer,J.,Dealing with Missing Data,in Research Letters in the Information and Mathematical Sciences,vol.3,PP.153 一 160,2002.

[12] Zaher Al Aghbari,Array-index:A Plug and Search K Nearest Neighbors Method for High-Dimensional Data,Data and Knowledge Engineering,PP.333-352,2005.

[13] Bruni Renato,Discrete Models for Data Impution,Discrete Applied Mathematics,vol.144,PP.59-69,2004.

[14] Quling Bo, Xiang bingren, An dengkui .Application of Artificial Neural Networks in medicine pattern recognition [J] Computers and Applied Chemistry, 2002, 04: 428-431.

[15] Qiao Yanjiang, Wang Xi, BI Kai-shun, Luo Xu.Application of Artificial Neural Networks in feature extraction of chemical pattern recognition of medicine Chansu[J] Pharmaceutical Journal, 1995,09: 698-701.

[16] Breiman L. Bag g ing Preditor s [J] . Machine Learning, 1996, 24(2) . Schapire R. E.The Strength of Weak Learn ability. Machine Learning.1990.

[17] Yoav Freund,and Robert E. Schapire.A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences.1997.

[18] Breiman L. Random Forests [J] . Machine Learning, 2001, 45(1) .