Fuzzy Systems and Data Mining VI
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA200713

Research on Characteristics of Chinese Herbal Medicine Compounds Based on Bisecting k-Means Algorithm

Yushu Wu [a*], Fenfen Xie [a,*], Lu Wang [a,1], Shoude Zhang [b], Lei Zhang [c] and Xiaoying Wang [a]
[a] *Department of Computer Technology and Applications, Qinghai University, Xining, 810016, China*
[b] *State Key Laboratory of Plateau Ecology and Agriculture, Qinghai University, Xining, 810016, China*
[c] *College of Computer Science, Information Management Center, Sichuan University, Chengdu 610065, China*

Abstract. The properties of Chinese Herbal Medicine (CHM) are determined to some extent by the properties of their molecular compounds, so it is of great significance to study CHM from the perspective of molecular compounds. In this paper, the clustering algorithm in data mining is used to study the relationship between the properties of CHM and its chemical components. Firstly, the molecular data are collected from the Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform, and the data set is preprocessed to extract the key molecular descriptors of chemical components. Secondly, the k-means algorithm and the Bisecting k-means algorithm are used to cluster the chemical components based on the CHM molecular descriptors, and the representative molecular features of the cold and hot CHM are extracted; finally, through experimental comparison, it is found that the clustering results obtained by Bisecting k-means algorithm are better. The clustering results show that the average values of molecular composition descriptors and charge descriptors in cold CHM are significantly higher than those in hot CHM. Therefore, the properties of CHM may be affected by molecular structure and molecular charge properties.

Keywords. bisecting k-means, Chinese Herbal Medicine, compounds, descriptors

1. Introduction

Chinese Herbal Medicine (CHM) has a long history of development. It is not only a bright pearl in Chinese culture, but also an important part of Chinese medicine and health. In recent years, the development of CHM has attracted more and more attention from all walks of life, and the research of CHM has also received widespread attention [1]. The properties of CHM are summarized by the predecessors through thousands of years of observation and long-term practice, and they are an important part of the

[*] These two authors contributed equally to this work.

[1] Corresponding Author, Department of Computer Technology and Applications, Qinghai University, Xining, 810016, China; E-mail: wlqgl@126.com.

whole CHM theoretical system. The four characters of CHM properties refer to the four properties of cold, hot, warm and cool [2-3]. For different patients with different diseases, traditional Chinese medicine practitioners often decide the CHM to be used by CHM properties [4]. Therefore, the study of the properties of CHM is of great significance for guiding the use of CHM.

In recent years, data mining technology has been widely used in the research of CHM properties. Liu Jingliang et al. used the ancient and modern literature as the basic data, and used the association rule algorithm to analyze the internal relationship between the properties and active ingredients of CHM. The research results show that there is a certain correlation between the efficacy of CHM and the properties [5]. Fu Xianjun et al. analyzed the molecular descriptors of the main components of CHM by using association rule algorithm, and found that the drug properties of CHM have a certain correlation with the molecular structure of the main components of CHM. The properties of CHM may be related to the molecular energy state of the main effective components of CHM [6]. Based on the records of "Chinese Pharmacopoeia", "Chinese Medicine", "Chinese Materia Medica Selected", "Chinese Medicine Pharmacology" and "New Chinese Medicine Journal", Wang Zhe analyzed the property, taste, meridian tropism, function and main treatment of CHM by frequency statistics and association rules, and found that "cold-bitter-lung-detoxification-sore throat" has the closest relationship [7].

However, up to now, there are still many problems and challenges in the research of CHM molecules. Firstly, there are few studies on the properties of CHM from a molecular perspective, and it is still in its infancy; secondly, the relationship between CHM properties and CHM molecules is not clear, and there is a serious disconnection between CHM and its chemical components. Thirdly, the traditional drug molecular research methods are tedious and time-consuming. Therefore, it is an urgent problem to explore new methods to explain the properties of CHM from the molecular perspective.

The purpose of this paper is to explore the relationship between the properties of CHM and its chemical components by data mining technology, and to find out the potential laws of properties. Firstly, the descriptors of CHM molecules were calculated, and then the binary k-means algorithm was used to cluster the CHM molecules. The results showed that the average values of component descriptors and charge descriptors in cold CHM were significantly higher than those in hot CHM. Therefore, the properties of CHM may be affected by molecular structure and molecular charge properties.

2. Research Methods

2.1. Molecular Descriptors

Molecular descriptors are a measure of the properties of a molecule in a certain aspect, and the molecular descriptors can describe the characteristics of a molecule from various aspects. Molecular descriptors can be roughly divided into: composition descriptors, topology descriptors, continuity descriptors, kappa descriptors, charge descriptors, molecular property descriptors, space descriptors, and so on. Because there are many kinds of molecular descriptors, this experiment screened out the molecular descriptors which have great influence on CHM properties in the data preprocessing stage [9].

2.2. K-means Clustering Algorithm

The K-means algorithm is an iterative algorithm. At the beginning of the algorithm, k-objects are randomly selected as clustering centers, and then the distance from each object to the cluster center is calculated, and the object is assigned to the nearest cluster center. A cluster center and objects belonging to it are collectively called a cluster. Every time an assignment is made, the cluster center will be recalculated according to the existing assignment objects, and the process of recalculation will continue until no objects are assigned to the new cluster, or the cluster center will not change, and the process will be terminated [10].

2.3. Bisecting k-means Algorithm

However, the k-means algorithm is sensitive to the initial clustering center, and the difference in clustering center often leads to a large volatility of clustering results [11].

To avoid that the initial clustering center will be on the same class, which will cause the algorithm to fall into the state of local optimal solution, this paper introduced another clustering algorithm: The Bisecting k-means Algorithm. The main idea of Bisecting k-means: first, all points are regarded as a cluster, and then the cluster is divided into two parts. Then a cluster with small sum of error squares is added to the cluster list, and the cluster with the largest sum of error squares is divided into two clusters [12]. This continues until the number of clusters is equal to the number k-given by the user. The square sum of clustering error is an index to measure the quality of clustering results. The smaller the value is, the closer the data points are to their centroids, indicating the better the clustering effect. So, the cluster with the largest sum of error squares should be divided first [13]. It can be seen that the main feature of this algorithm is to make up for the shortcomings of k-means algorithm. K-means algorithm cannot recalculate particles distribution between clusters very well when the clustering center is relatively far away, and this can give rise to a poor clustering result. The specific algorithm is described as follows:

Input: Data: Is data set; k: Is Number of clusters;
Output: L*owestSSE*: sum of the squares of the minimum errors;

```
1:      Select a cluster from the cluster table
2:      while centList < k          do\\when the number of clusters is less than k-
3:          LowestSSE → +∞\\Initialization minimum error
4:          for i = 0 → centList - 1    do
5:              Get data set samples belonging to cluster I
6:              k-means clustering of the cluster
7:              Get the sum of the errors after the cluster classi cation
8:              Get the sum of the error of the sample set that does not
                belong to the cluster
9:                  if sseSplit + sseNotSplit < LowestSSE        then
10:                     bestCentToSplit ← i\\Save best cluster
                        center
11:                     LowestSSE  sseSplit + sseNotSplit\\Save
                        the sum of the squares of the minimum
                        errors
12:                 end if
```

13: **end for**
14: *centList* ← *centList* + 1\\Number of clusters plus 1
15: **end while**

3. Data preprocessing

3.1. CHM Molecular Data Set

The data of this study comes from the Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform. 22 kinds of CHM were selected from the cold CHM in the database, and 33 kinds of CHM were selected from the hot CHM. Then 1419 different compounds were extracted from cold CHM, 2153 different compounds were extracted from hot CHM, 3199 compounds in total.

3.2. Calculation of Molecular Descriptors

We used the ChemoPy package of Python to calculate the molecular descriptors [14]. This paper focused on six kinds of descriptors, including molecular composition descriptors, topological structure descriptors, connectivity descriptors, kappa descriptors, charge descriptors and molecular property descriptors.

4. Experiment design

4.1. Algorithm Comparison

4.1.1. Comparison of Data Performance Indicators

The k-means algorithm and the Bisecting k-means algorithm were used to cluster all the data sets of cold CHMs. The distance formula was the European distance formula, and the number of clusters is 2~10 clusters.

Figure 1. The change of the sum of squared errors of two algorithms for cold CHM

It can be seen intuitively from Figure 1 that Bisecting k-means algorithm has a smaller sum of squared errors for each cluster than the k-means algorithm for the cold CHM data set. Therefore, Bisecting k-means algorithm was more suitable for cold CHM data set [15]. The line graph of the Bisecting k-means algorithm slowed down

significantly after the fourth cluster. Meanwhile, the k-means algorithm line graph slowed down at the fifth cluster. Therefore, 4 clusters were the best for the clustering result of Bisecting k-means, and 5 clusters were the best for the result of k-means algorithm.

Next, all data sets of hot CHM were clustered using the same way as clod CHM data.

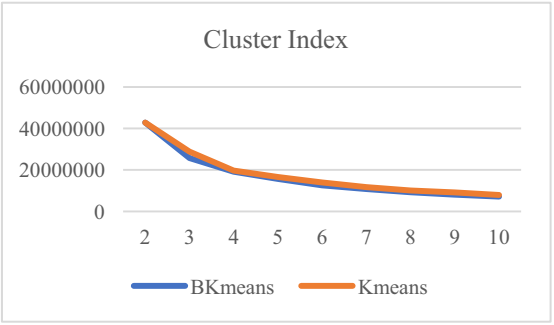


Figure 2. The change of the sum of squared errors of two algorithms for hot CHM

It can be seen from Figure 2, the Bisecting k-means algorithm was more suitable for hot CHMs. 3 clusters were the best for the clustering result of Bisecting k-means, meanwhile the result of k-means algorithm showed 4 clusters was the best.

4.1.2. Comparison of clustering results

As for the clustering of cold CHM data set, the fourth cluster result calculated by the Bisecting K-means algorithm and the third cluster result calculated by the K-means algorithm have the same number of compounds. The clustering results are shown in the following tables:

Table 1. The fourth cluster of compounds using the Bisecting k-means algorithm for cold CHM

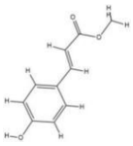
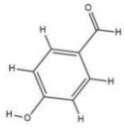
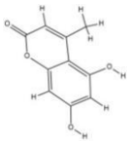
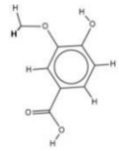
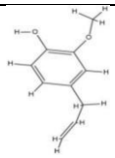
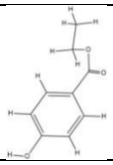
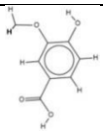
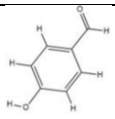
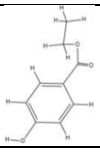
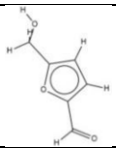
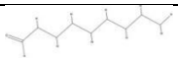
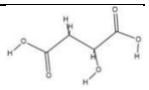
MOL001837	MOL001843	MOL001868	MOL000114	MOL000254	MOL001873
					

Table 2. The third cluster compounds by using k-means for cold CHM

MOL000114	MOL001843	MOL001873	MOL000748	MOL000116	MOL001468
					

Further observation of these two clusters of other elements showed that the element similarity of the fourth cluster calculated by the Bisecting k-means algorithm was very high, which basically has the characteristics of ring, double bond and oxygen atom, while the third cluster calculated by the k-means algorithm appeared the chain

like compound as shown in Figure 3. Thus, the clustering effect of the Bisecting k-means algorithm was better than that of the k-means algorithm [16].

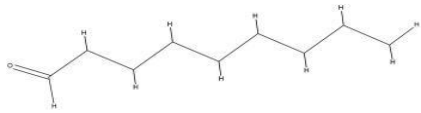


Figure 3. The third cluster compounds have low similarity with others calculated by the k-means algorithm

5. Clustering

5.1. Clustering Analysis

The representative molecules of each molecular descriptor in each cluster were found out through the median, which was calculated by the different molecular descriptors for each cluster. Additionally, if the value in the molecular descriptor is not the same as the median, then the numerator corresponding to the two closest values greater than and less than the median is taken. Finally, the sdf format of the compound was imported into the chemical draw software to visualize the compound [17].

Table 3. The representative compounds of each cluster in cold CHM calculated by Bisecting k-means

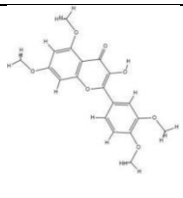
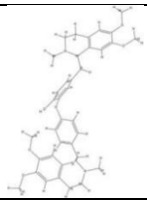
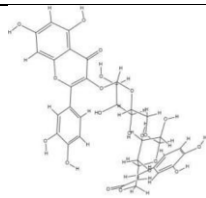
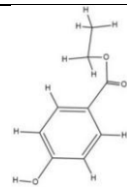
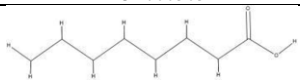
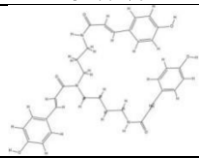
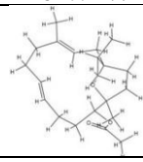
MOL008239	MOL001965	MOL003008	MOL001873
			

Table 4. The representative compounds of each cluster in cold CHM calculated by Bisecting k-means

MOL000303	MOL002764	MOL001278
		

As can be seen from the results, it turns out that the compounds in cold CHM mostly contained benzene rings. Additionally, the compound structure of cold CHM was more complex than hot CHM. Specifically, hot CHM have more chain structures.

The results of the parameters showed that the values of the important molecular descriptor parameters of cold CHM were almost higher than that of hot CHM.

Table 5. Comparison of the average value of molecular descriptor parameters of compounds in cold and hot CHM

Molecular descriptors	Cold property	Hot property
EWeight	500.4971602	375.7607915
nring	3.916804242	2.978081874
naccr	8.418577528	5.577138326
nsb	29.97761745	20.89608456
nta	73.03398545	56.35364337
TIAC	99.09199099	73.58934786

TPSA	145.1813007	91.48157419
FTPSA	145.1813007	91.48157419

Table 6. Comparison of the average value of the molecular descriptor parameters of each cluster of compounds in cold CHM

Molecular descriptors	Cluster 1	Cluster 2	Cluster 3	Cluster 4
EWeight	335.9508055	520.1070516	977.4764605	168.4543232
nring	2.484760522	4.351351351	7.710526316	1.120578778
naccr	3.956458636	8.071253071	19.51315789	2.133440514
nsb	17.93613933	30.14250614	64.17105263	7.660771704
nta	52.78955007	75.87714988	137.7763158	25.69292605
TIAC	65.49813788	101.8535971	195.7446711	33.27155788
TPSA	65.87849057	138.5108354	334.2089474	42.12692926
FTPSA	65.87849057	138.5108354	334.2089474	42.12692926

Table 7. Comparison of the average value of the molecular descriptor parameters of each cluster of compounds in hot CHM

Molecular descriptors	Cluster 1	Cluster 2	Cluster 3
EWeight	174.9042918	603.1325541	349.2455286
nring	1.23852459	5.026229508	2.669491525
naccr	1.106557377	11.82295082	3.80190678
nsb	9.577459016	33.60655738	19.50423729
nta	30.59344262	82.52786885	55.93961864
TIAC	34.20304877	117.7405213	68.82447352
TPSA	19.37276639	193.9616066	61.11034958
FTPSA	19.37276639	193.9616066	61.11034958

As can be seen from Table 6 and Table 7, it could be seen that the sorting results were basically the same if the value of each molecular descriptor is sorted between clusters, no matter it is a cold CHM or a hot CHM. For example, the numerical value of each molecular descriptor of the third cluster in cold drugs was basically the largest in each cluster, and the numerical value of each molecular descriptor of the second cluster in thermal drugs was basically the largest in each cluster.

Indicated from the above analysis results, the molecular weight-related descriptors have a great impact on the clustering results, and the average values of the molecular composition descriptors and charge descriptors in cold drugs are higher than those in hot drugs. It can be inferred that the medicinal properties of CHM may be affected by the charge properties and molecular structure.

5.2. Clustering discussion

There are some limitations of this research. The first is about the properties of molecular descriptors in the data set. The types of molecular descriptors in the data set used in this study are not comprehensive, and the number of compositions descriptors take up higher percentages, leading to the majority of conclusions is obtained from the point of view of atoms and molecular weight. Therefore, the conclusion is relatively one dimensional.

The second limitation refers to the algorithm used in this study. Although the Bisecting k-means algorithm overcomes the problem of the k-means algorithm converging to a local minimum, the algorithm considers that the attributes of the analyzed samples contribute uniformly to the clustering result, and does not consider the weighting of attributes. This may make the result less accurate [18-19].

6. Conclusions

This paper was based on the cluster analysis of molecular descriptors of CHM molecular compounds. Furthermore, it turns out that the Bisecting k-means algorithm could better seek and explore the characteristics of the compound in hot and cold CHM through the comparison of the application results of two clustering methods of k-means and Bisecting k-means. In addition, the algorithm found that EWeight (average molecular weight), nta (number of all atoms), TIAC (total information index of atomic composition), TPSA (Topological Polar Surface Area), FTPSA (Topological Surface Area Based on Fragmentation), the numerical values of compounds in cold CHM were significantly higher than those in hot CHM.

This research, where combined data mining technology and adopted new research ideas and methods to explore the medicinal properties of CHM, has great theoretical value and profound significance in promoting the modernization of CHM and the development of related industries.

There will be further improve for the algorithm, such as considering the contribution of attributes, and deeply discover the characteristics of cold and hot CHM. Finally, we still need to further explore the relationship and orderliness between the properties of CHM and chemical components.

Funding

This research was supported in part by the Applied Basic Research Programs of Science and Technology Department of Sichuan Province under grant number 2019YJ0110, in part by the Science and Technology Service Industry Demonstration Programs of Sichuan Province under grant number 2019GFW167, in part by the National Natural Science Foundation of China under grant number 61762074, in part by the Scientific Research Cooperation Project under grant number IPP-ZC-19071806, and in part by the Second Class Course Construction of Qinghai University in 2019 under grant number FL192002.

References

[1] Wu X. Data mining in the analysis of four properties for traditional Chinese medicine. Jinan University, Guangzhou, 2012;1-56.
[2] Chen L. Analysis on the theory of the rising and falling of traditional Chinese medicine. Guangxi Tradit. Chinese Med. 2020;43(02):44-7.
[3] Wang CY, Wang P, Wang ZG. Origin and development of Chinese medicine four-nature theory. Guangxi Tradit Chinese Med. 2009;33(01):8-10.
[4] Zheng BY. Study on the theory of traditional Chinese medicine based on its pharmacological action. *Electron J Clin Med Lit*. 2015;2(18):3718.
[5] Zhang LJ, Pei L, Li Y, Shi L. A study on association rules of Chinese herbal properties and the effective components in diuresis efficacy based on data mining. Chinese J Libr Inf Sci Tradit Chinese Med 2014;38(5):9-12.
[6] Fu XJ, Wang ZG, Li XB. Study on relationship between nature and multidimensional structure descriptor of main compounds from Chinese medicinal herbs. World Sci Technol Tradit Chinese Med 2017;19(04):549-55.
[7] Wang Z, Zhang PJ. Associated research on the traditional Chinese medicine nature. World Sci. Technol. Tradit. Chinese Med. 2013;30(08):859-63.

[8] Chen ZK, Song X, Gao J, Zhang JL, Li P. Research progress in the data mining-based TCM diagnoses. Chinese J. Tradit. Chinese Med. 2020;1-15.
[9] Cao DS. Chemopy: A software package of QSAR / QSPR molecular descriptor calculation based on Python. China Chemical Society. Abstracts of papers of the 11th National Conference of Computer Chemistry. China Chemical Society: China Chemical Society 2011;1.
[10] Kant S, Ansari IA. An improved k-means clustering with Atk-inson index to classify liver patient dataset. Int J Syst Assur Eng Manag. 2016;7(1).
[11] Wu XB, Guo Q, Zhang LB, Liang YZ, Liu JG. Multidimensional data clustering based on network community detection method. Int J Syst Assur Eng Manag. 2020;37(02):421-3.
[12] Liu GC, Huang TT, Chen HN. Improved Bisecting K-means clustering algorithm. Comput. Appl Res. 2015;32(02):261-3.
[13] Wang JY, Wan QYn, Yan TW. Improved Bisecting k-means algorithm based on hadoop. Comput. Appl Softw. 2016;(9):4-8.
[14] Zhang L. Study on the drug-likeness of Chinese herbal medicine based on molecular descriptors and application. Harbin University of Commerce, Harbin, 2019;62.
[15] Li YK. Analysis and comparison of three typical clustering algorithms in data mining. Comput Knowl Technol. 2020;16(15):52-6.
[16] Yang L, Peng LF, Feng BH. Component analysis of traditional Chinese medicine based on Data Mining—Application of cluster analysis in the determination of ginsenosides in Panax ginseng by HPLC. Softw Guid. 2009;(08):169-171.
[17] Peng T, Sun LY, Zhou JJ. Hierarchical clustering of traditional Chinese medicine database based on Molecular Fingerprint. Softw Guid. 2013;30(06):575-581.
[18] Liu WW. The research on varying weight k-means and its application. Xiamen University, Xiamen, 2017;(07):64.
[19] Wu B. Improvement and application of correlation weighted k-means algorithm. Jiangxi University of Science and Technology, Nanchang, 2018;(07):58.