# Feature Extraction of Chinese Materia Medica Fingerprint Based on Star Plot Representation of Multivariate Data

CUI Jian-xin[1, 2], HONG Wen-xue[1], ZHOU Rong-juan[3], GAO Hai-bo[1]

1. College of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China
2. Key Laboratory of Measurement Technology and Instrumentation of Hebei Province, Qinhuangdao 066004, China
3. Maternal and Child Health Hospital of Qinhuangdao, Qinhuangdao 066004, China

**Abstract**: **Objective** To study a novel feature extraction method of Chinese materia medica (CMM) fingerprint. **Methods** On the basis of the radar graphical presentation theory of multivariate, the radar map was used to figure the non-map parameters of the CMM fingerprint, then to extract the map features and to propose the feature fusion. **Results** Better performance was achieved when using this method to test data. **Conclusion** This shows that the feature extraction based on radar chart presentation can mine the valuable features that facilitate the identification of Chinese medicine.

**Key words**: Chinese materia medica; feature extraction; fingerprint; multivariate graph; radar chart presentation

**DOI:** 10.3969/j.issn.1674-6384.2011.02.009

## Introduction

The key point of the qualitative analysis of Chinese materia medica (CMM) is to build up a scientific, advanced, and feasible fingerprint diagram technique. The CMM fingerprint diagram is a visual presentation of physical and chemical information of Chinese medicines, which indicates their substance base through spectroscopic or chromatographic analysis (Zou and Yan, 2008; Shao, 2009). And it is chemical chromatography curve, which is characterized by similarity with the overall and obscurity with the individual, meaning with the fingerprint. At present, the Chinese fingerprint technology has been applied to the quality appraisal of the Chinese medicines, Chinese herbal medicines (CHM), and other natural herbal medicines, and evaluation of stability between or among the batches. The core of the method is to build the CMM fingerprints, to extract the characteristics of the CMM fingerprint using information processing, and to determine the authenticity of CHM, the stability of its quality, and the possessions or batches of drugs according to these characteristics. However, the most of the researchers in this field are of the chemical analysis and pharmaceutical, whose study focused on the acquisition and the establishment methods of fingerprinting (Zhu and Wang, 2005; Luo, Wang, and Cao, 2000). The majority of the researchers use the chemical methods to analyze the diagram, of which the similarity is observed only by the number of fingerprints peak, peak position, peak of the order, and the ratio between the peak and the peak area ratio (Pan, Wang, and Ye, 1994; Liu *et al*, 2005). So, it can be said that the study of the fingerprint stays in the establishment of testing methods and the correlation between physical indicators (Dai *et al*, 2007; Zhao *et al*, 2004). Therefore, the study of CMM fingerprint needs to introduce mathematical information processing method to expand the obtained fingerprint data processing method in depth. Recently, this problem has attracted concerns of the researchers in signal and information processing, principal component analysis, cluster analysis, fuzzy model identification, and wavelet transform, and other mathematical methods in medicine fingerprint analysis have been applied.

It has been discovered that every one of these methods can not guarantee the integrity of map data or poor visual and difficult to understand. In addition, the chromatographic retentions in the chromatogram peaks are vulnerable to a variety of factors, and it is difficult to analyze the fingerprint. To resolve this issue, data

First author: Cui JX Tel: +86-335-8074 715 E-mail: ydcuijianxin@yahoo.com.cn
Received: July 28, 2010; Revised: October 29, 2010; Accepted: November 8, 2010
Fund: the National Nature Science Foundation of China (60873121)

needs to be processed from a large number of equipments, to extract information on chemical characteristics, so as to realize pattern recognition to enable them to evaluate and control the quality of CHM and their preparations. In this paper, aiming at the above problems, a novel feature extraction method of the CMM fingerprint based on the principle of radar chart was advanced. The quantitative radar map characteristics of the CMM fingerprint features have strong intuition, which will help to understand and to ensure the unity of integrity and obscurity effectively. Attention is concerned on the radar chart presentation of the CMM fingerprint data, the feature extraction and fusion of the radar graph and the construction of standard CMM fingerprint. The advantage of this method is the visualization of the CMM data, the feature extraction process, and the classification process and performance. Experimental results show that the method has achieved better classification performance.

**Presentation and process of CMM fingerprint data based on radar chart**

In the traditional method, a multi-dimensional data sample is regarded as a point in multi-dimensional space. So, from a geometric point of view, it is the non-chart feature. The relationship between the dots describes the relationship between the samples, such as distance relationship, but does not reveal the internal data structure of the sample points directly. This method has a visible problem that the data with more than 3-dimensional can not be visualized. The graph representation of multivariate data not only illustrates the internal data structure of this sample well, but also depicts the relationship among samples. In other words, using the map to figure the non-map feature may form the distinctive graphic features which are conductive to the visual classification or clustering.

**Principle of radar chart presentaion**

The characteristics of the CMM fingerprint data are small sample and many variables. Typically, only 2- or 3-dimensional Cartesian coordinates can be expressed, while the radar chart can present three to dozens of dimensional data. The traditional radar mapping steps are as follows. Firstly, making a circle, and the circle is divided into $p$ equal parts. Secondly, linking the center of the circle and these $p$ points in turn, and the $p$ radiuses are defined as the axis of each index and marked with the appropriate scale. Finally, the $p$ index values of a given observation are collected in the corresponding axis, and then link them into a p-gon. This p-gon may reflect the relationship between multi-variable data of an observation. Moreover, the p-gon has a new feature, named as graphic feature, which is different from the traditional characteristics. And it includes physical characteristics, mathematical characteristics, and structural features. The graphic features of multi-variable contain more extensive data structure information, which can be divided into local features and overall features. The overall features include area, orientation, location, and center of gravity vector. The local features include the adjacent amplitude ratio, geographical area ratio, and symmetry, *etc*. For example, if the shapes of the polygons are different, or the direction of the maximum variables, or the area of the polygons, it means that the sort from whom the radar chart presents is different. Using the graphic features of multi-variable we will probably get better classification results.

As the CMM fingerprint data associated with the time sequence, this paper presents a special form of radar chart. We define the coordinates of the outer circle as the time coordinate (relative $t_R$), the highest point of the circle as zero and it is also the last relative $t_R$ of the mapping data, clockwise as positive, the coordinates of the radius as radial coordinates, the direction from the center of the circle to the circumference as positive. The p-gon got in this way preserves well the integrity of fingerprint data. At the same time, the radar chart presentation and feature extraction on a large number of data reduce the impacts of various factors on the fingerprint data, as well as get a very good visual graph. The radar chart (Fig. 1) is formed by the relative $t_R$ and peak area. And the data are from the "Traditional Chinese Medicine and Digitization" table 5-11 (Zou and Yan, 2008), take the front seven peaks of the sample 1 for example.

**Feature fusion of radar chart**

Usually fingerprint data, the parameters of the CMM fingerprint data sample include the relative retention value (i.e. the relative $t_R$), peak value, peak area, $n$ strong peaks, characteristic peak number, total peak number in common, and the overlap rate. These parameters are mapped into radar chart, which is used

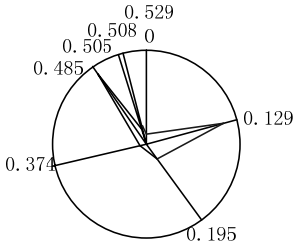


**Fig. 1 Radar chart of relative $t_R$ and peak area**

to describe the whole fingerprint data in the feature space. We often get three or more radar charts and a higher dimension, which are not conductive to visibility, comprehension, and classification. In this paper, we proposed the feature fusion of the radar chart. Firstly, the non-equidistant radar charts were drawn. Secondly, the graphic features were extracted and made the input feature of the next level, so the equidistant radar charts were obtained. Thirdly, the graphic features on the equidistant radar charts were extracted until the number of the feature met the classification requirements. Finally, a radar chart with more variables is used to describe the entire fingerprint (Fig. 2) (Meng *et al*, 2008).

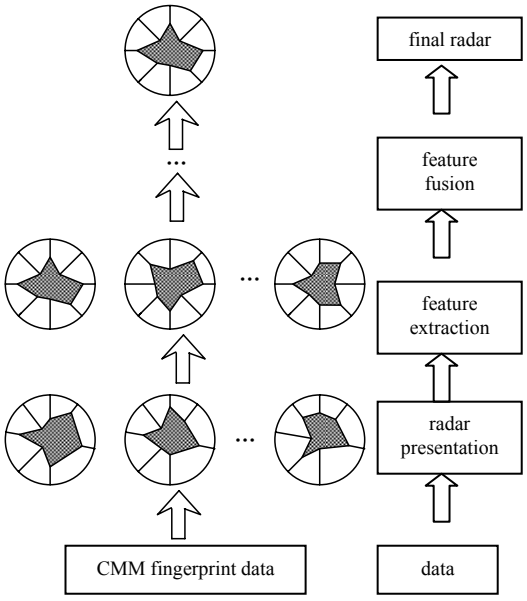


**Fig. 2 Feature fusion process of radar chart**

**Radar chart center of gravity**

Science (Tang and Guo, 2001) and Nature (Liu *et al*, 2006) published respectively the study that the drosophila recognized different objects through the center of gravity height, direction, and other parameters and form memories, which showed that the graph center of gravity might be an important feature beneficial to pattern recognition. From the geometric perspective, the radar graph of sample data forms a polygon and produces a center of gravity, which constitutes to the center of gravity graphic feature.

Easy to know, the dimension of the radar chart's center of gravity with $m$ variables is $C_n^m$, if the sample has $n$ variables, and the total dimension of the sample is $C_n^2 + C_n^3 + \cdots + C_n^n = (2^n - n - 1)$. There are two specifical situations. One is the adjacent two variables constitute a triangle. From the geometric perspective, three midlines of a triangle intersect in one point, the intersection is called the center of gravity of the triangle. And the distance between the center of gravity and the vertices of the triangle is equal two times of the distance between and midpoint of the subtense. The other is all the variables form the center of gravity of n-gon, and obviously the dimension of the radar chart's center of gravity with $n$ variables is 1. It contains the global information of samples. The characteristics of the radar chart include the vector modulus of center of gravity (the amplitude of the origin to the center of gravity) and direction. The following introduction takes an example of the radar chart's center of gravity with two variables, and other situations are similar.

The feature extraction of the radar chart's center of gravity with eight variables is shown in the chart (Fig. 3). The adjacent variables $x_1$ and $x_2$ constitute the triangle AOB and express the center as G, and the distance from the origin O to the center of gravity G is expressed as OG. Then, the feature extraction formula for the local center of gravity the adjacent two variables constitute in radar chart is as follows (Gao *et al*, 2007).

$$\begin{cases} abs_i = \sqrt{(\frac{r_i + r_{i+1}\cos\omega_i}{3})^2 + (\frac{r_{i+1}\sin\omega_i}{3})^2} \\ angle_i = arctg(\frac{r_{i+1}\sin\omega_i}{r_i + r_{i+1}\cos\omega_i}) \end{cases}, i = 1, \cdots, d$$

Here, $abs_i$ and $angle_i$ express respectively the amplitude of the center of gravity and the true angle of the triangle, the $i$-dimensional variable and the $i + 1$ dimensional variable constitute. While $r_i$ and $r_{i+1}$ express the value of the $i$-dimensional variable and the $i + 1$ dimensional variable after normalization is generally calculated by the pretreatment. The radian $\omega_i$ is the curvature between the $i$-dimensional variable and the $i + 1$ dimensional variable. When the circle is divided equally by dimension $d$, it is $\omega_i = 2\pi/d$.

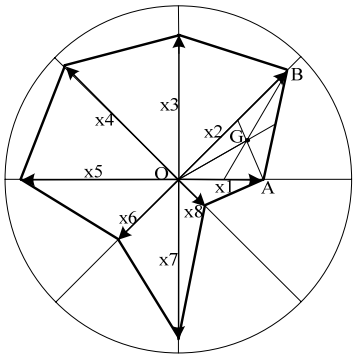


**Fig. 3 Feature extraction of radar chart center of gravity with eight variables**

## Experiment

The experimental data are from the "Traditional Chinese Medicine and Digitization". The data set includes *Cassia obtusefolia* L., *C. tora* L., and *C. occidentalis* L., it has 42 samples and the sort is known. First of all, the data set is mapped into radar chart (non-equidistant), then extracts the graphic features from the non-equidistant radar chart, makes it the input feature of the next level, and finally gets a radar map with variable more. In the case of the sample large enough in size, the radar chart can also be a standard fingerprint describing some CMM.

Linear classifier (lc), quadratic classifier (qdc), k nearest neighbor classifier (knnc), and parzen classifier (parzenc) were used. The classification features include the center of gravity vector, area, direction, and symmetry. The contrast features are the original features and the principal component features. The testing evaluation indicators are the errors of the 10% samples tested by 100 cross validation. The result is shown in Table 1. The error rate (lc, qdc, knnc, parzenc, and average error) of the radar chart features was significantly lower than that of the original features and the principal component features. This shows that the feature extraction based on radar chart presentation can mine the valuable features that facilitate the identification of Chinese medicine.

**Table 1 Classification error rate**

| Original feature | Error rate / % | | | | |
|---|---|---|---|---|---|
| | ldc | qdc | knnc | parzenc | average |
| Original | 0.1020 | 0.0902 | 0.0933 | 0.0876 | 0.0933 |
| PCA | 0.0867 | 0.0733 | 0.0661 | 0.0764 | 0.0333 |
| Radar chart | 0.0100 | 0.0133 | 0.0200 | 0.0231 | 0.0166 |

## Conclusion

Combining with hard- and soft-threshold, the research shows that the feature extraction and the standard fingerprint's construction technique of the CMM fingerprint data, on the basis of the radar graphical presentation theory of multivariate, are the potential method. It incarnates a novel thought on the CMM fingerprint pattern recognition, and puts forward a new method on the feature extraction of the CMM fingerprint. The data experiment expresses that we obtained better performance on the radar chart over the traditional features. At the same time, it possesses the characteristic of visualization, and it is propitious to the human-computer interaction of the study process and the comprehension of the problem.

### References

Dai Y, Ye YQ, Feng WB, Wang H, 2007. Application of pattern recognition in quality control of the traditional Chinese medicine. *J Yunnan Nationalities Univ* 16(4): 334-337.

Gao HB, Hong WX, Cui JX, Xu YH, 2007. Optimization of principle component analysis in feature extraction. IEEE Press: *Int Conference on Mechatronics and Automation (ICMA2007)*, 3728-3732.

Liu G, Seiler H, Wen A, Zars T, Ito K, Wolf R, Heisenberg M, Liu L, 2006. Distinct memory traces for two visual features in the Drosophila brain. *Nature* 439(7076): 551-556

Liu SH, Zhang XG, Zhou Q, Sun SQ, 2005. Use of FTIR and pattern recognition to determine geographical origins of Chinese medical herbs. *Spectrosc Spectral Anal* 25(6): 878-881.

Luo GA, Wang YM, Cao J, 2000. The characteristic fingerprint of multi-dimensional and multi-data and its application. *Chin Tradit Pat Med* 22(6): 395-397.

Meng H, Hong WX, Song JL, Wang JJ, 2008. Analysis of protaein mass spectra based on multivariate feature fusion visualization. *J Yanshan Univ* 32(5): 451-456.

Pan XL, Wang W, Ye CY, 1994. General paragon of the application of fuzzy clustering analysis in the classification of atractylodes. *J China Pharm Univ* 25(6): 348-352.

Shao JQ, 2009. Advances in studies on fingerprint of Chinese materia medica. *Chin Traidit Herb Drugs* 40(6): 994-998.

Tang S, Guo A, 2001. Choice behavior of Drosophila facing contradictory visual cues. *Science* 294: 1543-1547

Zhao Y, Liang YZ, Yi LZ, Xie PS, Yang H, 2004, Application of chemical pattern into recognition of Zhishi (*Fructus Aurantii Immaturus*). *Res Pract Chin Med* 18: 70-73.

Zhu EY, Wang XR, 2005. An orthogonal expansion algorithm of principal component suitable to deal with the fingerprinting data of Chinese medicine. *J Xiamen Univ*: *Nat Sci* 44(6): 884-885.

Zou CC, Yan HY, 2008. *Traditional Chinese Medicine and Digitization*. Anhui Science and Technology Publishing Company: Hefei, 21.