Report on fingerprints classification

1.Problem Introduction

Human fingerprints are unique, and almost no human fingerprints are exactly the same. Therefore, fingerprints are usually used as a human identification feature. Furthermore, we often call the features that can be used to uniquely identify things as fingerprint features. As an attempt to study fingerprints, we decided to study the classification of fingerprints.

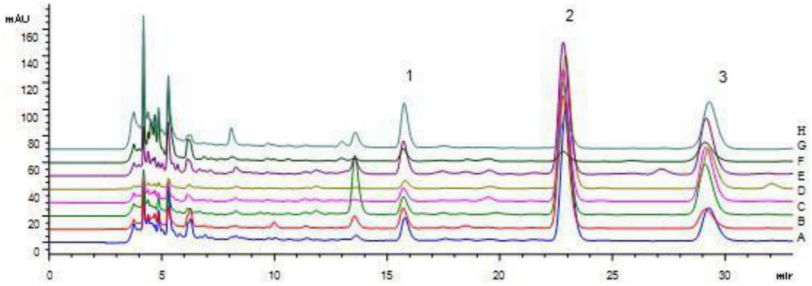


图3 不同来源蜂胶HPLC指纹图谱
A：内蒙古；B：辽宁；C：山西；D：河北；E：山东；F：安徽；G：浙江；H：福建

Fingerprints refer to the chromatograms or spectra of some complex substances, such as certain tissues or cells, and proteins that are properly processed and analyzed using certain analytical methods to indicate their chemical characteristics. Fingerprints are mainly divided into traditional Chinese medicine fingerprints, DNA fingerprints and peptide fingerprints. In many fields of scientific research, when people can't fully understand every minute structure of the research object, they turn to analyze it as a whole. Fingerprint recognition technology, as one of the biometric identification technologies, has gradually matured in the new century and entered the field of human production and life.

We hope that we can infer which type of substance the spectrum belongs to from the fingerprint spectrum of the unknown type, so as to achieve the effect of classification and identification. The problem we want to solve is to train our neural network model using samples of known classifications, and realize the classification of fingerprints of

unknown classification in the first step. At this stage we can classify samples of unknown classification into known classification. Considering that there are many fingerprints in nature that may not belong to any of the known classifications, in order to better solve these new class samples, in the second stage, we introduced a spatial geometric classification method, which assists classification by three-dimensional distance.

2. Formalization

There are many ways to construct classification methods, such as extracting certain characteristics of fingerprint maps, giving their mathematical representations: elements of geometric space or vector space, etc., and then selecting or constructing classification methods suitable for this mathematical representation; another example is constructing probability Statistical model, then use statistical methods to classify and so on.

We preliminarily decided to analyze the geometric features of the fingerprint and determine the feature points to be selected through the image and extract them through the neural network. Finally, through the establishment of a three-dimensional coordinate system, the use of geometric distance to achieve the classification of fingerprints.

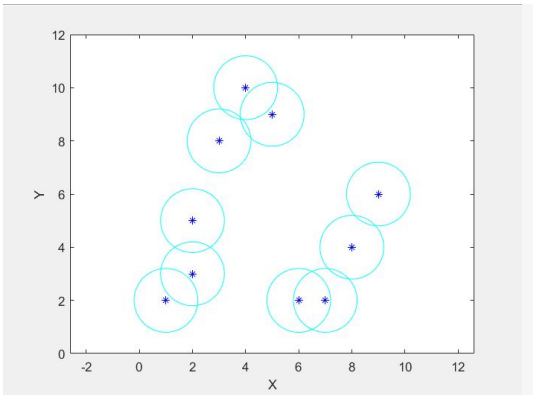
Fingerprint lines are not continuous, smooth and straight, but often interrupted, bifurcated or turned. These breakpoints, bifurcation points and turning points are called "feature points". The parameters of the feature point include direction (the node can face a certain direction), curvature (describe the speed at which the grain direction changes), and position (the position of the node is described by x/y coordinates, which can be absolute or relative to a triangle Point or feature point).

We select three features according to the trend of fingerprint atlas, use the nine samples of known classification given in the title to train the initially established neural network, and try to optimize the initial parameters through Backpropagation until the input features are processed by the neural network The error between the

final characteristic value and the expected result is minimized. Next, input the five samples that need to be classified into the neural network, compare the output features with the three known features of A, B, and C respectively, and classify the samples into the known class that is closest to its feature value.

So as to complete the classification. Aiming at the expansion of the model, we apply the model from the supervised learning into an unsupervised learning process. Use the trained neural network to classify the classification samples that does not belong to the known type by using its weight, that is, perform feature extraction on the fingerprint maps of 14 other categories, and then process these features in our ANN model to attain the three eigenvalues of each sample. We can regard the three eigenvalues of each sample as the coordinate of a point in a 3D space. The position of these points in the 3D space stands for the relationship of these samples, which means we can realize how different two samples are by comparing the distance between them. Then put the sample points that are close to each other into the same cluster, to accomplish the clustering process of these unknow samples.

For the clustering method, we choose an innovative way based on the hierarchical clustering, Spherical Area Overlapping Method. The implementation of this method is to let each sample point be the center of a sphere with an increasing radius. When increase the radius gradually, the space area inside the sphere will also increase and these spheres will overlap with each other. When two spheres are overlapping, their center points are clustered into the same group. After all spheres have overlapped with others, this algorithm can be finished, we put the points that are overlapped together into a cluster.

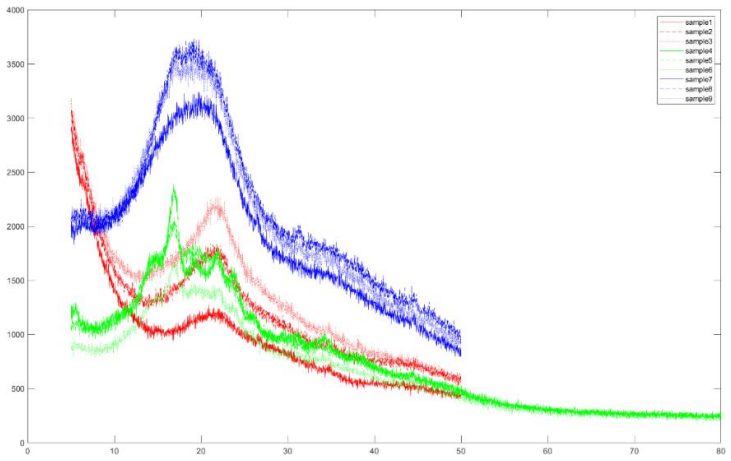


3.Algorithm

The first thing we need to do is to extract features from the 1-9 sample data that have been classified into ABC categories, and then use the remaining five samples to test the classification effect. We can easily find that this type of problem belongs to supervised learning, that is, the process of using samples of known categories to adjust the parameters of the classification model to achieve the purpose of classifying unknown samples. Therefore, we need to use the parameters of known samples to find the commonality between samples of the same category and the differences between samples of different categories, and then use these commonalities and differences to build a classification model. When the unfamiliar samples are substituted into the model, the ones that are shown as common are the same category, and the ones that show the difference as different categories, and then the classification of the unfamiliar samples is realized.

If you want to describe an image more comprehensively, just one feature value is not enough. You need to extract image features from different angles. The more angles, the more accurate the description of the image and the higher the degree of distinction between it and other images. Therefore, we choose the neural network model.

The neural network model takes each type of feature value of a sample as input data, which is located in the Input Layer. Therefore, it is first necessary to draw the image of the sample and extract its features. Use MATLAB software to draw the images of samples 1-9, as shown in the figure below.



According to the image, we can see the significant difference of ABC image:

1. At the end point, the B type abscissa value reaches about 80, and the AC type abscissa value is about 50.

2. Considering that the maximum value of A is obtained near the starting point, in order to obtain the "peak" size of each image, that is, the maximum point that the image can reach after an upward trend, 500 points from the starting point (roughly corresponding to the image The abscissa is equal to the maximum value of the data after 10) as the "peak value".

It can be concluded that the "peak value" of category C is above 2500 and category AB is below 2500

3. At the starting point, the ordinate value of category A is the largest, above 2500; the value of category BC is smaller, below 2500.

Therefore, three characteristic values can be extracted: the abscissa at the end point, the maximum value after the 500th point (hereinafter referred to as the maximum value), and the ordinate at the starting point. After normalizing the features, the data obtained are as follows:

```
[ 2.2490e+01,  6.5664e-01,  3.9746e+00]
[ 2.2490e+01,  1.2209e+00,  3.5988e+00]
[ 2.2490e+01,  1.5819e+00,  3.0566e+00]
[ 3.7490e+01,  3.2261e+00,  7.6041e-01]
[ 3.7490e+01,  2.6450e+00,  5.5257e-01]
[ 3.7490e+01,  2.6725e+00,  6.9409e-01]
[ 2.2490e+01,  1.9636e+00, -2.8068e-05]
[ 2.2490e+01,  2.1312e+00, -8.8898e-02]
[ 2.2490e+01,  2.0388e+00, -9.8136e-02]
[ 2.2490e+01,  2.4910e+00,  2.2855e-01]
[ 2.2490e+01,  1.5901e+00,  1.8285e+00]
[ 3.7490e+01,  3.1585e+00,  6.9991e-01]
[ 3.7490e+01,  2.3661e+00,  8.3152e-01]
[ 2.2490e+01,  2.1071e+00, -1.8039e-01]
```

Then there are the classification criteria for each feature: whether the end point abscissa value is greater than 65 (the average of 50 and 80), whether the maximum value is greater than 2500, and whether the starting point ordinate is greater than 2500.

Next, initialize the parameters of the model. Although the weight parameters can be automatically set up by pytorch, we need to design the input, output layers as well as learning rate to make sure the better performance of classification.

```
self.fc1 = nn.Linear(in_features=3,out_features=3)
self.out = nn.Linear(in_features=3,out_features=3)
```

```
optimizer = torch.optim.Adam(model.parameters(),lr = 0.1)
```

Considering that there are three features, each feature can distinguish one type, so 0 and 1 can be used to represent the classification result of the feature, for example: in the classification of feature 1, if the result is 1, it is judged as class B, if the result is 0, it is judged as type A or type C, and the other types are the same. In this way, the classification results are combined into a column vector. After three categories of judgment, we can accurately classify the sample categories. The detailed results are as follows.

1-D	0	1	2
3-D	[1,0,0]	[0,1,0]	[0,0,1]
CLASS	B	C	A

Use Forward Propagation Algorithm to verify the reasonableness of weights:

The input function formula of each node of the hidden layer:

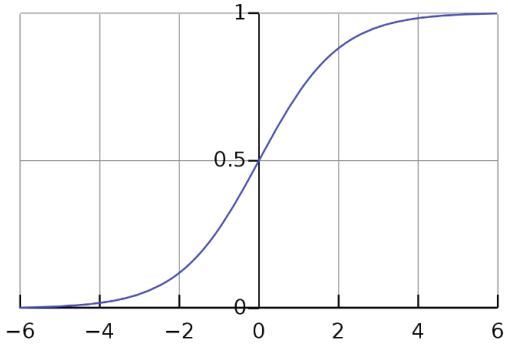
$$I = w_0X_0 + w_1X_1 + w_2X_2 + w_3X_3$$

Substituting data to get the input value V, substituting V into the excitation function to get the output value O, the formula is as follows:

$$O = g(I)$$

Special note: The sigmoid function is selected as the excitation function in this model, and the expression and function image are as follows:

$$g(I) = \frac{1}{1 + e^{-I}}$$



$I < 0 \;\; g(I) < 0.5, I > 0 \;\; g(I) > 0.5$

Passing the value Q obtained by multiplying the output value O of each node by the weight w to the output layer, and combine the three Qs into a 3×1 matrix $\begin{bmatrix} Q_1 \\ Q_2 \\ Q_3 \end{bmatrix}$, and then round each element of the matrix to get the final classification result.

In torch, we need to transform the 3D array to 1D array otherwise the program cannot recognize the label. The program is as follows:

```
train_y = [[0, 0, 1],
           [0, 0, 1],
           [0, 0, 1],
           [1, 0, 0],
           [1, 0, 0],
           [1, 0, 0],
           [0, 1, 0],
           [0, 1, 0],
           [0, 1, 0]]
train_y= np.argmax(train_y, axis=1)
train_y = torch.LongTensor(train_y)
train_y
```

```
tensor([2, 2, 2, 0, 0, 0, 1, 1, 1])
```

We have known that$X_1 < 65, X_2 < 2500, X_3 > 2500$， so$I_1 > 0, I_2 > 0, I_3 < 0$. After deal by sigmoid function, we get $O_1 > 0.5, O_2 > 0.5, O_3 < 0.5$， passing to the output layer $Q_1 > 0.5, Q_2 > 0.5, Q_3 < 0.5$ .After rounding, the classification result is $\begin{bmatrix}1\\1\\0\end{bmatrix}$, in line with expectations. Similarly, get B label $\begin{bmatrix}0\\1\\1\end{bmatrix}$, C label $\begin{bmatrix}1\\0\\1\end{bmatrix}$. This shows that the estimated value of the weight is more reasonable.

Use the backpropagation algorithm to update the weight value to find the most suitable weight:

In order to improve the classification accuracy, it is necessary to find a more appropriate weight, which uses the back propagation algorithm.

We should understand that the error between the data before rounding and the standard result is caused by the weight, and the process of reducing the error is the process of optimizing the weight. Therefore, there must be a connection between the error and the weight. We try to find the functional relationship between the error and the weight, that is, the cost function, and then use the gradient descent algorithm to find the target weight value when the error is the smallest.

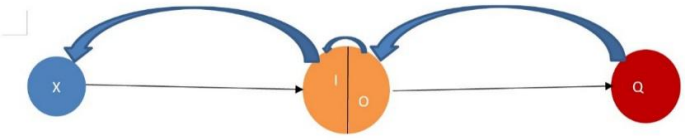
First, define the cost function:

$$E = \frac{1}{2}(Q - S)^2$$

Q is the actual value; S is the standard value. Then, use the chain rule to establish the derivative relationship between the cost function and the weight of each layer：

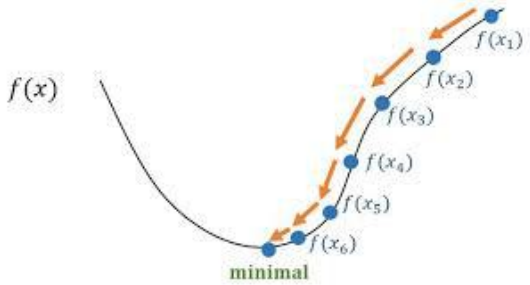
$$First\ layer：\ \frac{\partial E}{\partial w_1}=\frac{\partial E}{\partial Q}\times\frac{\partial Q}{\partial O}\times\frac{\partial O}{\partial I}\times\frac{\partial I}{\partial w}=(Q-S)\times w_2\times O\times(1-O)\times X$$

$$Second\ layer：\ \frac{\partial E}{\partial w_2}=(Q-S)\times O$$



Update weight value：

$$w^{'}=w-\alpha\times\frac{\partial E}{\partial w}(\alpha\ is\ the\ learning\ rate)$$



The upper procedure can be programmed like this:

```
optimizer.zero_grad()
loss.backward()
optimizer.step()
```

So far, the weight has undergone an update, but to achieve the expected effect, it is necessary to iterate continuously and update the weight value continuously, so we set up the epochs as 1000 so that the error can converge effectively.

Expand this model to solve an unsupervised problem:

Now we have built an ANN model by the training samples, it can also be used to clustering some unlabeled sample into different groups.

We apply Spherical Area Overlapping Method for clustering, let all the sphere have the same radius. In order to identify whether two spheres are overlapped, we can see whether the sum of their radius is greater than the distance between their center:

$$R_1 + R_2 > d_{12}$$

The key to the problem is to find the optimal value of R. If R is too small, the classification cannot be completed; if R is too large, the number of classes will be too small, and the two points with large difference may be classified into one cluster.

In order to solve this problem, we define a function:

$$N = func(R)$$

, where N is the number of clusters and R is the radius of the sphere. Plot the function by MATLAB, to visualize their relationship. The function consists of many horizontal line segments. The length of the line segment can be used as the basis for determining the optimal radius R. The longer the line segment is, the longer the duration of maintaining the number of categories N at a fixed value when R is increased, and the classification result is more stable. Therefore, the most reliable value of N is corresponding to the longest line segment (the longest range of R). We can find the optimal radius that has a constant cluster number lasting for a longest span. Then use this radius to finish the final clustering.

4.Experiment Result

Next is our specific solution process:

1. Substitute the training samples into the model again, and get the results of the final feature values of nine samples:

```
[3.7472e-03,  4.2142e-11,  9.9625e-01]
[3.8491e-03,  4.4586e-11,  9.9615e-01]
[4.5346e-03,  6.2931e-11,  9.9547e-01]
[9.9391e-01,  3.4398e-03,  2.6512e-03]
[9.9083e-01,  7.9428e-03,  1.2298e-03]
[9.9003e-01,  8.8240e-04,  9.0856e-03]
[4.6952e-03,  9.9530e-01,  5.4666e-10]
[4.0727e-03,  9.9593e-01,  4.1632e-10]
[4.1306e-03,  9.9587e-01,  4.2773e-10]
```

fig.1

After rounding, the sample classification result is:

[2, 2, 2, 0, 0, 0, 1, 1, 1]

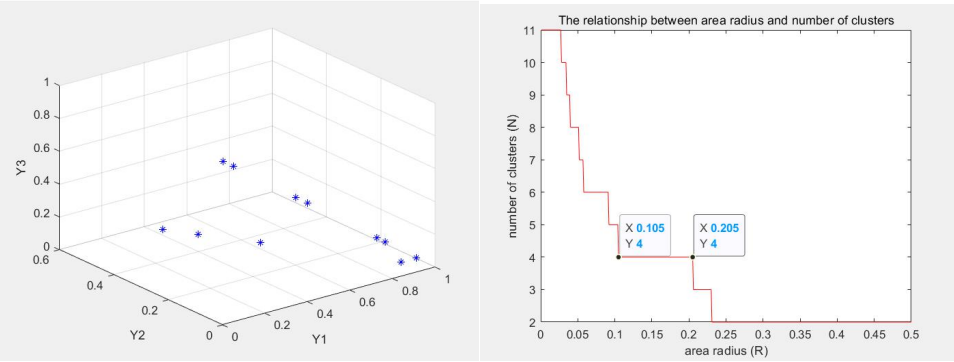
The result meets the known conditions of the title, indicating that the model training is successful.

4. Substitute 5 test samples for classification, and get the result:

[1, 2, 0, 0, 1]

the predicted result is highly close to the real result, and the algorithm can be considered to perform well.

5. Clustering the test sample points and visualize the relation between N and R:



From the figure above, we can find that: when 0.105＜R<0.205, the line segment is longest, the clustering result is the most stable. At this time, the clusters number N = 4. Then use the radius within this range to execute the algorithm of Spherical Area Overlapping Method, we can attain the result below:

```
>> Clustering_3
r = 0.15
Test1 is belong to cluster1
Test2 is belong to cluster2
Test3 is belong to cluster1
Test4 is belong to cluster2
Test5 is belong to cluster3
Test6 is belong to cluster4
Test7 is belong to cluster2
Test8 is belong to cluster3
Test9 is belong to cluster4
Test10 is belong to cluster4
Test11 is belong to cluster1
Test12 is belong to cluster4
Test13 is belong to cluster4
Test14 is belong to cluster2
```

5.Discussion

Now that we achieve the supervised learning by means of ANN, the basic concept is to abstract the features of figures and set up the expected values corresponding different labels, then feed the data to the model to cater the labels. It is remarkable that the output value like fig.1 is the combination of features, they somehow represent the comprehensive features of a certain figure.

Why we cannot use ANN model to preprocessing the features and then use some clustering algorithms to achieve unsupervised learning classification of unlabeled objects? Note that the combination of features data is more precise and smarter in classification than the traditional isolated features data. Thus, the next step is to combine the ANN model and clustering algorithm to achieve unsupervised learning with a set of unlabeled data.

Summarize the advantages of our model and algorithm. We use neural network and other algorithms as the classification method. Through the steps of extracting features and training parameters, the features of each sample are three-dimensionally classified and classified from the perspective of geometric distance, which gives good results for classification. In the expansion, the stability of the radius increase is analyzed, and a relatively stable radius increase is selected.

At the same time, we believe that our model and algorithm can be further improved in the future, such as:

1. Three feature points are selected when selecting feature points. If more feature points with resolution can be selected, it can be applied to the classification of more unknown category data.

2. In the future, the number of hidden layers can be appropriately increased, and the accuracy of the model can be improved by increasing the number of hidden layers before reaching overfitting.

3. In the classification of the second question, the abscissa span of one sample is

obviously different from other samples, and the processed feature value cannot distinguish it separately.

4. In the process of selecting the learning rate, the method is awkward, and the parameters are determined only through trial and error, and the steps are complicated and lengthy. We believe that another model can be introduced to optimize the learning rate, thereby improving the intelligence and efficiency of the model.