上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

COURSE PAPER

Title: Evolution and analysis of American music's genre

Course :	Basic Mathematical Statistics	
Member 1 :	秦子健	021020910010
Member 2 :	徐德全	021020910068
Member 3 :	杨新宇	021020910058
Member 4 :	赵学涛	021010910018

Contents

Chapter 1 杨新宇的部分 1
Chapter 2 Influence Among Artists 2
2.1 A directed network of influencers and followers 2
2.2 Similarity of music features among artists 2
2.2.1 Euclidean similarity 2
2.2.2 Cosine similarity 4
2.2.3 Similarity of music characteristics among artists 4
2.3 A comparison between two influence networks 4
Chapter 3 Cluster Analysis 7
3.1 Data Preprocessing 7
3.1.1 Entropy Weight Process 7
3.1.2 PCA Dimensionality Reduction 7
3.2 Cluster Method 7
3.2.1 K-means Cluster 7
3.2.2 Hierarchical Cluster 8
3.3 Cluster Evaluation 8
3.3.1 Within-cluster Sum of Squared Errors(SSE) 8
3.3.2 Silhouette Analysis 8
3.4 Visualization of Clusters 8
Chapter 4 Machine Learning 9
4.1 Neural network structure of PNN 9
4.1.1 Radial Basis Function(RBF) 9
4.1.2 Bayesian optimal decision theory 10
4.2 Summary of implementing PNN 11

Chapter 1　杨新宇的部分

Chapter 2 Influence Among Artists

Some artists can list a dozen or more other artists who they say influenced their own musical work. It has also been suggested that influence can be measured by the degree of similarity between song characteristics, such as structure, rhythm, or lyrics. In this chapter, a directed network of influencers and followers is established according to "influence_data" reported by the artists themselves, as well as the opinions of industry experts. In addition, the similarity of song characteristics among different artists is analysed to reflect the influence in another way. After a brief comparison between two influence networks, it may lead to some interesting results.

2.1 A directed network of influencers and followers

According to "influence_data" and preceding entropy weight analysis, a directed network of influencers and followers can be established with each artist having a weight coefficient. For convenience of analysis and visualization, a sub-network including top 100 influential artists is built and plotted in Figure 2–1.

In Figure 2–1, the size of a node represents the influence score of an artist and the direction of an arrow is from a follower to an influencer. It is clear that most of the top 100 artists are from genre Pop/Rock (in red colour). The Beatles, Bob Dylan, and The Rolling Stones come top 3 influential artists, which consists well with the previous analysis and the reality. The Beatles is regarded as the most influential band of all time; Bob Dylan is one of the greatest songwriters, and The Rolling Stone is one of the most famous rock bands. It is noticeable that in this network, the link between two artists has a specific direction, that is to say, one is the influencer while the other is influenced, which is different from similarity analysis. Moreover, an influencer can also be a follower, giving the fact that the most influential artist The Beatles is also influenced by "The Band" and Bob Dylan. This reveals the mutual influence of artists on each other.

2.2 Similarity of music features among artists

Similarity between song characteristics can also reflect musical influence among artists. To be clear, two artists sharing the similar music style would have a greater chance to be influenced by each other. This section is aimed to build another influence network based on similarity analysis.

2.2.1 Euclidean similarity

In $\mathbb{R}^n$, the Euclidean distance $||x - y||_2$ between two vectors $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ is always defined. It corresponds to the L_2-norm $||.||_2$ of the difference x–y between the two vectors. It can be computed as:

$$\|x - y\|_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} \tag{2–1}$$

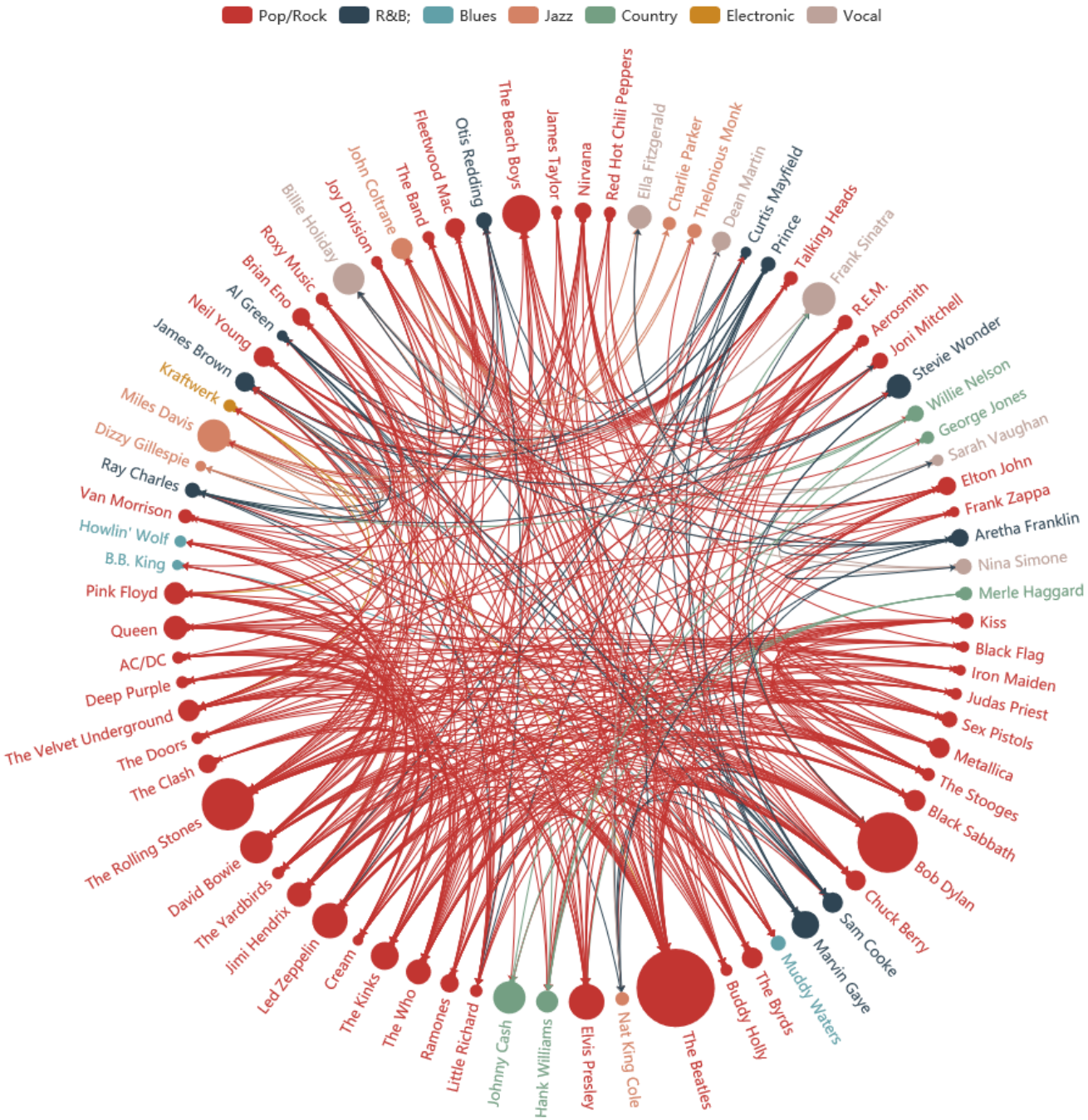


Figure 2–1 A directed network of influencers and followers

Here, PC1–PC7 are 7 major components reflecting the music style, so the Euclidean distance between two artists' music characteristics is:

$$dist_{i,j} = \sqrt{\sum_{k=1}^{7} \left(PC_k^i - PC_k^j\right)^2} \tag{2–2}$$

In order to get a similarity coefficient ranging from 0 to 1, we define Euclidean similarity as:

$$Sim_{L2}^{i,j} = \frac{1}{dist_{i,j} + 1} \tag{2–3}$$

2.2.2 Cosine similarity

In data analysis, cosine similarity is a measure of similarity between two sequences of numbers. For defining it, the sequences are viewed as vectors in an inner product space, and the cosine similarity is defined as the cosine of the angle between them, that is, the dot product of the vectors divided by the product of their lengths, as shown in Equation (2–4). It follows that the cosine similarity does not depend on the magnitudes of the vectors, but only on their angle. The cosine similarity always belongs to the interval [-1,1].

$$Sim_{cos}^{i,j} = \frac{\sum_{k=1}^{7} PC_k^i PC_k^j}{\sqrt{\sum_{k=1}^{7} \left(PC_k^i\right)^7} \sqrt{\sum_{k=1}^{7} \left(PC_k^j\right)^2}} \tag{2–4}$$

2.2.3 Similarity of music characteristics among artists

The data set "data_by_artis" provides us with more than 10 musical features of each artist, such as danceability, tempo, loudness and etc. After PCA and entropy weight analysis, we select 100 most influential artists to measure their similarities among each other. Euclidean similarity could reflect the distance between 2 vectors in the space, while some problems may occur when two vectors are close to the original point but in totally different directions. Cosine similarity is useful to reveal the similarity in directions, neglecting the information of length. It would be more reasonable to consider both Euclidean similarity and cosine similarity in this case. We define that two artists share the similar music style (have a potential to influence each other), when:

$$Sim_{cos}^{i,j} > 0.7 \qquad \text{and} \qquad Sim_{L2}^{i,j} > 0.7 \tag{2–5}$$

Under such criterion, a new influence network is established based on similarity analysis, as shown in Figure 2–2.

2.3 A comparison between two influence networks

So far, two influence networks have been obtained. The first directed one (Figure 2–1) is reported by the artists themselves, as well as the opinions of industry experts. The latter one without arrows is established by the degree of similarity between song characteristics, such as structure, rhythm, or lyrics. We are curious about the difference between them, and hopefully we expect to find some interesting information.

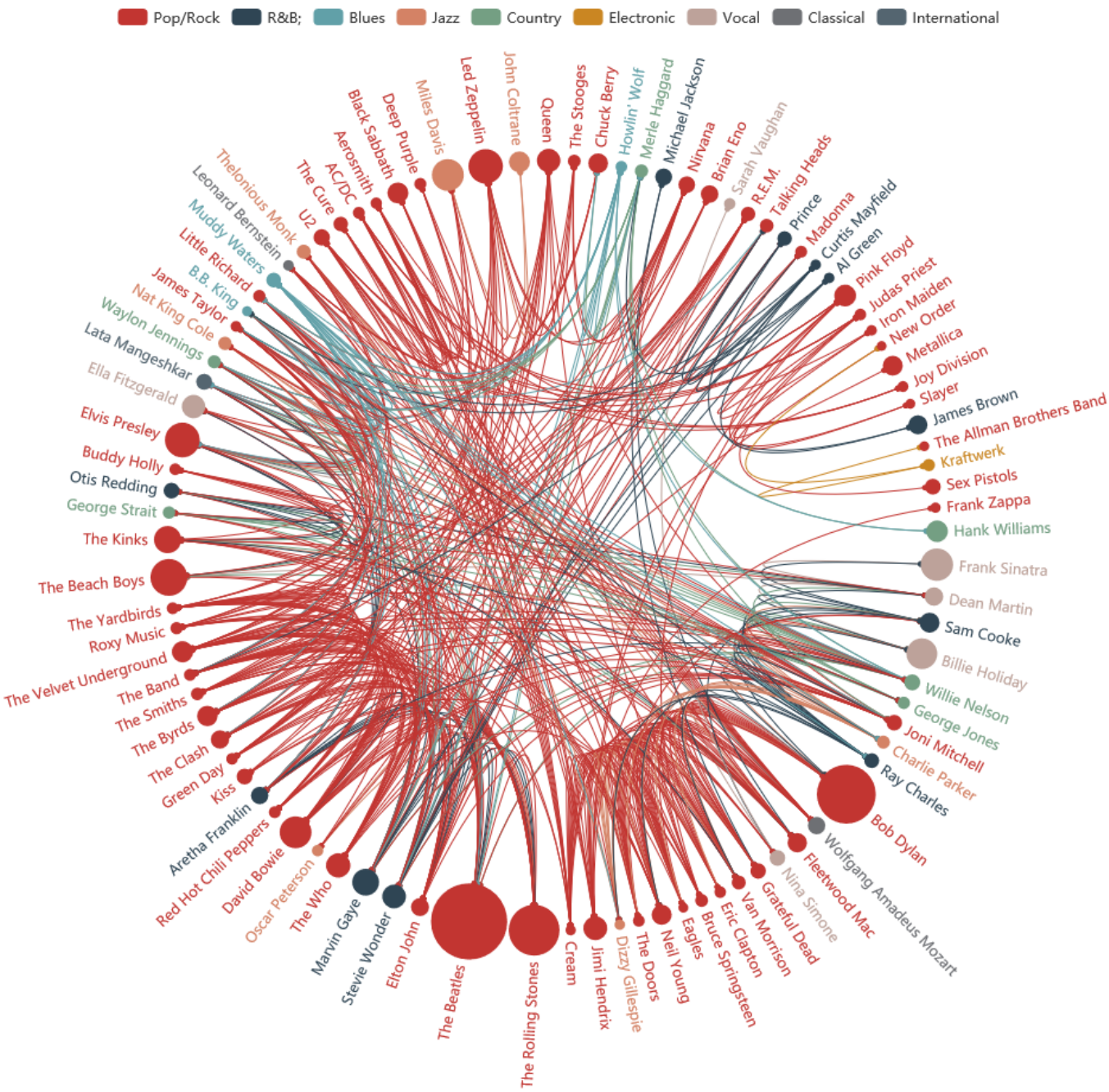
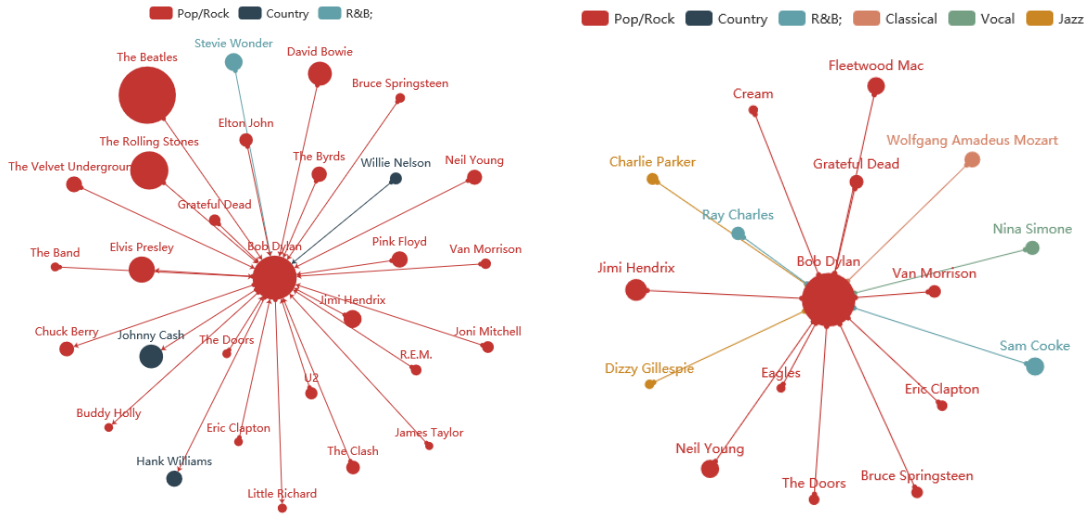


Figure 2–2 An influence network based on musical similarity

Generally speaking, most artists are mainly influenced by artists within the same genre as reported in the first network, while more between-genre-influence relationships are witnessed in the second network. This phenomenon is contributed to the subjectivity involved in the interviews or comments. Artists tend to claim they follow the famous musicians in the genre they belong to. In addition, industry experts often start from a professional manner to make judgements, that is to say, the genre barrier cannot be overlooked in this scenario.

To be specific, each artist has a list of "nominal" (self-claimed) followers and a list of potential professional fans (based on musical similarity). These two lists may overlap to some extend, indicating that a "double-checked" strong bond among artists exists.

For simplicity and convenience, the influence data of the famous artist Bob Dylan is taken as an example for analysis. Two mini influence networks related to Bob Dylan are shown in Figure 2–3a and 2–3b.



a) reported by artists & experts

b) computed from similarity analysis

Figure 2–3 Bob Dylan influence networks

The left figure shows the artists that have influenced on or been influenced by Bob Dylan according to data reported by artists & experts. The right one illustrates the artists sharing similar music features with Bob Dylan. 7 artists appear in both networks, indicating a strong influence — not only orally judged by person, but also reflected in their musical works. They are: Grateful Dead,Van Morrison, Eric Clapton, Bruce Springsteen, Neil Young, The Doors and Jimi Hendrix. We also checked whether some self-claimed followers have a very different music style with their influencer, and the result shows that no one does.

Chapter 3 Cluster Analysis

In the original data, we divided the artists into genres, but the music created by one person may have different styles, so we need to cluster the music data and classify it. In the clustering method, two methods, K-means clustering and hierarchical clustering, are selected, and the clustering results are evaluated.

3.1 Data Preprocessing

Before performing cluster analysis, the data needs to be preprocessed. The data is processed into multi-dimensional vectors, and each vector can be regarded as a sample point, so as to calculate the distance between the sample points, and then determine the ownership of each sample point.

3.1.1 Entropy Weight Process

First of all, in the selection of data, since the original data contains nearly 100,000 songs, we only need to analyze the songs created by influential artists, so we select the top 100 influential artists determined by the entropy weight method. He has composed more than 20,000 songs in total. In terms of specific data processing, the list of the top 100 influential artists has been determined by the entropy weight method in the second chapter above, and pandas is used to select the exported table and extract all the songs created by the corresponding artist.

3.1.2 PCA Dimensionality Reduction

The features of music are more complicated, and it is not easy to extract features, so when clustering, we use PCA to reduce the data dimension from 12 dimensions to 7 dimensions, and process the results. When clustering, the weights in PCA are also added. The processing of the weight part of PCA has been completed in the first chapter, we know the PC value and weight of each feature, and remap the data from 12 dimensions to 7 dimensions

3.2 Cluster Method

Mainly using K-means and hierarchical clustering methods

3.2.1 K-means Cluster

3.2.1.1 Principle and Process

The basic idea of K-means clustering is to find a partitioning scheme of K clusters (Clusters) iteratively, so that the loss function corresponding to the clustering results is minimized. Among them, the loss function can be defined as the sum of squared errors of each sample from the center

point of the cluster to which it belongs:

$$J(c,\mu) = \sum_{i=1}^{M} \left\| x_i - \mu_{c_i} \right\|^2 \tag{3–1}$$

Where x_i represents the ith sample, c_i is the cluster to which x_i belongs, μ_{c_i} represents the center point corresponding to the cluster, and M is the total number of samples.

The core goal of K-Means is to divide a given dataset into K clusters (K is a hyperparameter) and give the center point corresponding to each sample data. The specific steps are very simple and can be divided into 4 steps:

1. Data preprocessing. Mainly standardization and outlier filtering.
2. Select K centers randomly, denoted as $\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_k^{(0)}$
3. Define loss function: $J(c,\mu) = \sum_{i=1}^{M} \left\| x_i - \mu_{c_i} \right\|^2$
4. Let t=0,1,2,... be the number of iteration steps, and repeat the following process until J converges:
   (a) For each sample x_i, assign it to the nearest center:

$$c_i^t <- \operatorname{argmin}_k \left\| x_i - \mu_k^t \right\|^2 \tag{3–2}$$

   (b) For each class center k, recalculate the class center:

$$\mu_k^{(t+1)} <- \operatorname{argmin}_\mu \sum_{i:c_i^t=k}^{b} \|x_i - \mu\|^2 \tag{3–3}$$

3.2.1.2 Advantages and Disadvantages

3.2.2 Hierarchical Cluster

3.2.2.1 Principle and Process

3.2.2.2 Advantages and Disadvantages

3.3 Cluster Evaluation

3.3.1 Within-cluster Sum of Squared Errors(SSE)

3.3.2 Silhouette Analysis

3.4 Visualization of Clusters

Chapter 4 Machine Learning

We apply PNN(Probabilistic Neural Network) for machine learning.PNN can be regarded as a Radial Basis Neutral Network,it's based on the Radial Basis Function(RBF) network and involved Density function estimation and Bayesian decision theory.Under most conditions,PNN can realize the discriminant boundary asymptotically approaching the Bayesian optimal decision surface.

4.1 Neural network structure of PNN

PNN is consist of input layer,Radial Basis Layer(hidden layer) and competitive layer(output layer),which is shown in 4–1.The first layer is the input layer,it's used to receive training sample and transfer data to hidden layer.The hidden layer is Radial basis layer,it receives samples form the input layer and return a scalar which depends on the Euclidean distance to the center vector.the scalars are transferred to the output layer,which is a competition layer,which can output the scalar that satisfy the competition requirement.During the training process,we are able to achieve the center vector and variance of each class.In the later testing process,through calculating the Euclidean distance from the sample to each center vector and transform into the probability of decision based on Bayesian optimal decision.

4.1.1 Radial Basis Function(RBF)

The main idea of the RBF layer is using the RBF as the basis of the latent space,this allows the input vector to be directly mapped to the latent space without connecting through weights.When the center point of the RBF is determined after the training process,the mapping relationship is also determined.Among them, the role of the hidden layer is to map the vector from the low-dimensional `p` to the high-dimensional `h`, so that the low-dimensional linear inseparability can become linearly separable to the high-dimensional,which is the main idea of the kernel function. In this way, the mapping of the network from input to output is nonlinear, while the network output is linear with respect to adjustable parameters. The weights of the network can be directly solved by the linear

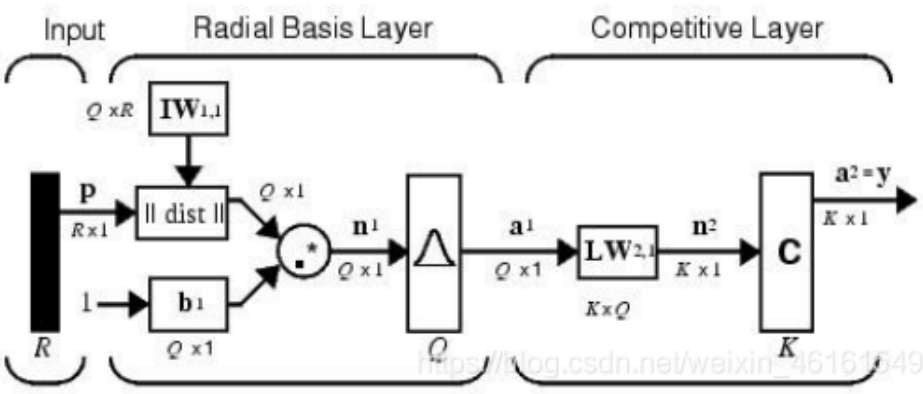


Figure 4–1 structure of PNN

equation system, which greatly speeds up the learning speed and avoids the local minimal problem. the activative function of the raidal basis neural network can be written as

$$R(x_p - c_i) = exp(-\frac{1}{2\sigma_i^2}\|x_p - c_i\|_2^2) \tag{4–1}$$

As we can see,if the sample vector x_p is closed to the center vector c_i,R would be more closed to 1,based on this,we are able to build the connction from R to the probability of decision through Bayesian optimal decision.

4.1.2 Bayesian optimal decision theory

The competition layer(output layer)is based on the Bayesian optimal decision theory. for a typicalbinary classification problem: $c = c_1$ or $c = c_2$.Prior probability can be written as

$$h_1 = p(c_1), h_2 = p(c_2), h_1 + h_2 = 1 \tag{4–2}$$

for any input vector $x = [x_1, \ldots, x_n]$,the classification criterion are shown below:

$$c = \begin{cases} c_1, p(c_1|x) > p(c_2|x), \\ c_2, otherwise \end{cases} \tag{4–3}$$

$p(c_i|x)$ is defined as when x happens,the posterior probability of c_i.According to the Bayesian function.the posterior probability equals to

$$p(c_i|x) = \frac{p(c_i)p(x|c_i)}{p(x)} \tag{4–4}$$

during the decision making process,the classification criterion should be the posterior probability.In practical situation,cost of misclassification are need to be considered.The cost of missclassify type 1 sample into type 2 and missclassify type2 sample into type1 are sometimes different.Therefore,the classification are needed to be adujusted.Define event α_1 is classify the input vector into type c_i,the cost of misclassification is λ_{ij}.So the expectation cost of apply event α_1 is

$$R(\alpha_i|x) = \sum_{j=1}^{N} \lambda_{ij} p(c_i|x) \tag{4–5}$$

Assume that the cost of correctly classification is 0,the expectation cost of event α_1 is

$$R(\alpha_i|x) = \lambda_{12} p(c_2|x) \tag{4–6}$$

The Bayesian theorem turns into

$$c = \begin{cases} c_1, R(c_1|x) < R(c_2|x), \\ c_2, otherwise \end{cases} \tag{4–7}$$

However,in the project,we assume that the cost of missclassification are equals,so the classification only depends on the probability of events.

4.2 Summary of implementing PNN

Firstly,based on the label of the training set,we are able to attain the center vector u_j and variance σ_j of class j.

$$u_j = min \sum_i \|x_{ij} - u_j\|_2^2 \tag{4–8}$$

$$\sigma_j = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \|x_{ij} - u_j\|_2^2} \tag{4–9}$$

where K is the number of class j.

Secondly,for each new sample x_i in the testing layer,we apply the RBF to transform the Euclidean distance between x_i and center vector of each classes u_j into a real number ϕ_{ij} ranging from 0 to 1.then,we normalize ϕ_{ij} and regard it as the probability of classification.

$$\phi_{ij} = exp(-\frac{1}{2\sigma_j^2}\|x_{ij} - u_j\|_2^2) \tag{4–10}$$

$$p_{ij} = \frac{\phi_{ij}}{\sum_j \phi_{ij}} \tag{4–11}$$

which means the probability of classifying sample x_i into class j.

Lately,we apply the Bayesian optimal theorem as the decision criterion and output the result.

In order to clearify and simplify the result,we only select several singer's songs as the entire sample.randomly choose 80% of the entire sample(16891) as the training set while the rest 20%(4222) is the testing set.the Accuracy of the training set and the testing set are around 97%(shown in 4–2),which indicates the effectiveness and accuracy of the PNN to the problem.

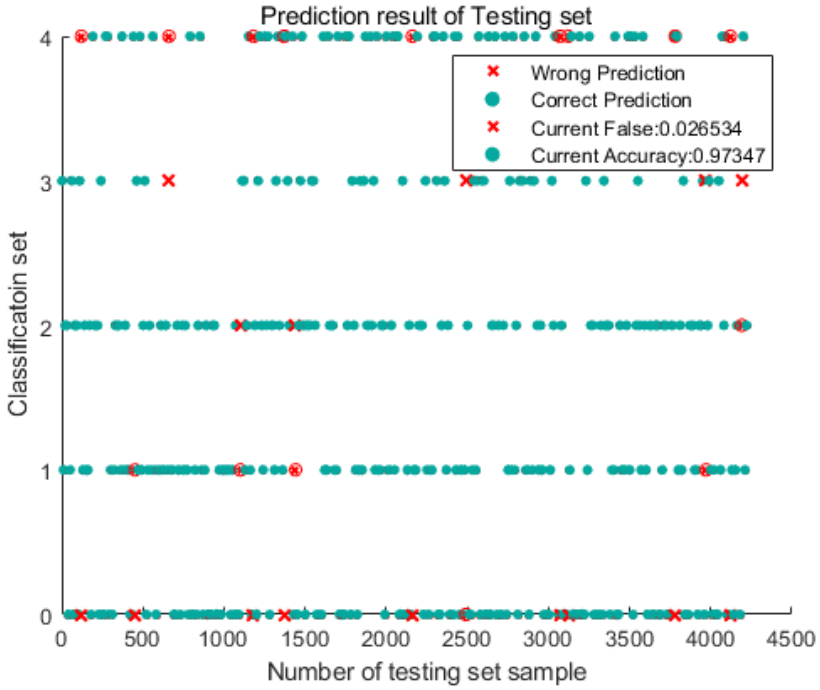


Figure 4–2 Prediction result of Testing Set