The lab report of the Ali music Forecast

Zhang Junwei
eMail: 474650206@qq.com

Jan 30th, 2018

1 Introduction

After 7 years of development and precipitation, Ali music has millions of music library resources, tens of millions of users are active on the platform everyday, and there are hundreds of millions of users' auditions and collections.In the original artists and works, but also has tens of thousands of independent musicians, upload every month tens of thousands of original works, formed over hundreds of thousands of tracks of the original works, grasp the data repository for such a huge music trend has a very important guiding role.

This practice is based on the music playback data Ali user's history to mini the trend of the artist on Ali music platform.Thus realizing the music trend in a period of time the accurate control.In our experiments, we first preprocessed the data and counted the total amount of songs played by each artiste. We used matplotlib.pyplot to display artistes' playback with images, effectively predicting the trend of artists' play.

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data.Models for time series data can have many forms and represent different stochastic processes. When modeling variations in the level of a process, three broad classes of practical importance are the autoregressive (AR) models, the integrated (I) models, and the moving average (MA) models.

These three classes depend linearly on previous data points.[25] Combinations of these ideas produce auto regressive moving average (ARMA) and auto regressive integrated moving average (ARIMA) models.

2 Experiment

2.1 Data Sets

The data includes two files, the user's behavior table and the song artist table. User behavior tables include user-id, song-id,gmt-create, action-type (1. playback, 2. download, 3. collection), and Ds(Recording time). The song-artists file includes: song-id, artist-id, publish-time, song-init-plays, language and gender.

2.2 Data preprocessing

According to the final evaluation method of the evaluation, we need to calculate the amount of play per artist in the prediction time, so I get the daily playback volume of each artifact through feature extraction and data integration.

We get the amount of data for each file. The song data file includes 100 artists, 26958 songs, 9 languages, 3 kinds of members, and user behavior data files, including 349946 users, 3 behavior types (play, collection, download), etc.The two files are shared by the song ID, so the two files are merged with the song ID.

data	number
artists	100
songs	26958
users	349946
languages	9
members	3
user actions	3

2.3 Modeling

The ARIMA method is used in this experiment. 1.Obtain the time series data of the observed system.

2.For data drawing, the observation is a stationary time series; for the nonstationary time series, the d order difference operation is first carried out to be transformed into a stationary time series.

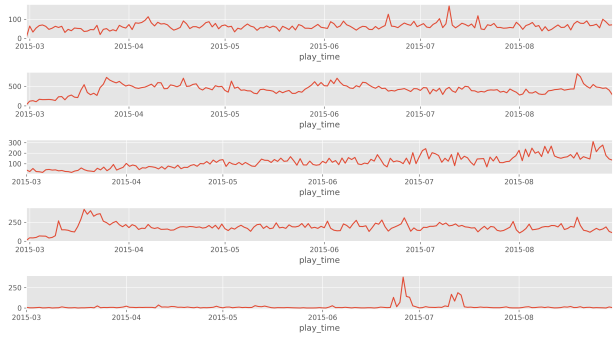

Figure 1: Three images of 6-10 artist play.

3.After second steps, the stationary time series has been obtained. The auto correlation coefficient ACF and partial autocorrelation coefficient PACF of stationary time series are obtained respectively, and the best class p and order q are obtained by analyzing auto correlation map and partial auto correlation graph.

But in our lab, I don't use the ACF and PACF, I just try to adjust the parameters in our model. For different artist, the parameters are sometimes different. So in the experiment, I user the anomaly detection to find the best parameters for the artist.

From the player's trend chart, we can see that most of the artists play a stable trend, but there will be occasional sudden increase or sudden drop. So we can reduce the impact of sudden increase or sudden drop through the log function.

4.The ARIMA model is obtained by the d, q and p obtained above. And then start the model test of the obtained model

Since the time series of the whole is decomposed before, we need to decompose the previous season and rejoin the result after the prediction.

ARIMA model

Input:Training time series data set, the test series data set
Output:Parameter, p, d, q(class p, order a, d order difference operation)
Initialization:converts the playback matrix to DataFrame
repeat:predict every artist
Calculation:the value of F

2.4 Evaluation

That artist j number T_{jk} play, in the actual k days, the contestants set for U, the artist collection is W, the contestants i artist j on day k to play the number of $S_{i,j,k}$, the contestants for artists to play j predicted and actual variance normalized variance $\sigma_{i,j}$

$$\sigma_{i,j} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\left(S_{i,j,k} - T_{j,k}\right)T_{j,k}\right)^2}$$

The weight of the artist j is the square root of the entertainer's playback

$$\phi_j = \sqrt{\sum_{k=1}^{N} T_{j,k}}$$

Contestant i's prediction is F_i

$$\sum_{jew}(1 - \sigma_{i,j}) * \phi_j$$

At the beginning, we calculated the F value of a medium value playback matrix, which is 5700,Then we calculate a new F value by the time series, which is 5400. The effect is not as good as the simple median method, because the time series is only effective for some artists. The data contain some very low data artists, or even play 0.Therefore, I combine the median

and time series models, set the initial value to the median, when the playback volume is less than 0, the initial value is the median, and then use the time series model to process the data.

The experimental results show that the size of the F value is closely related to the initial median value. The more the initial value is closer to the player's true playback, the higher the F value, the lower the lower.

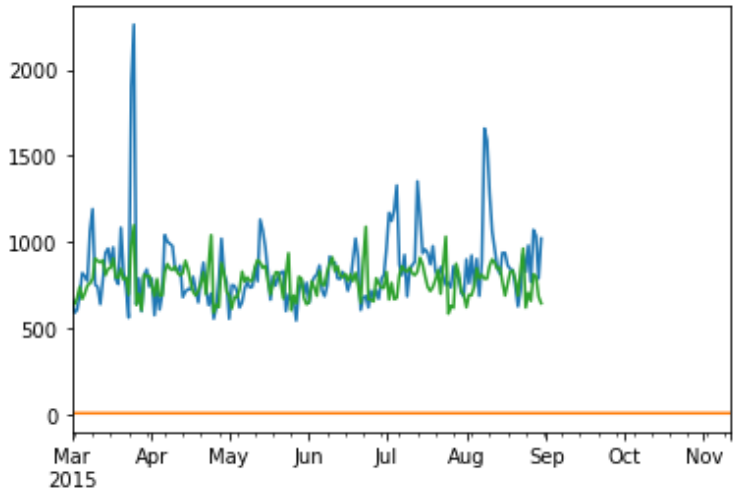

Figure 2: Three result of predict play.

3 Summary

Through this experiment, I first understand the time series analysis model and have a good understanding of data mining. In addition, I have a deeper understanding of the importance of data storage structure, and how to select a convenient and efficient data storage structure, which not only improves efficiency in data processing, but also facilitates data manipulation.

Besides, for a model, it is very important to choose a model parameter suitable for data, which requires us to have a comprehensive understanding of the trend and structure of data.