阿里音乐流行趋势预测

李艳东（2120151006），曹文强（2120150977），樊荣 （2120150982），张川（2120151059）

I 开题

A. 问题描述

阿里音乐拥有数百万的曲库资源，拥有数亿人次的用户试听、收藏等行为。在原创艺人和作品方面，拥有数万的独立音乐人，每月上传上万个原创作品，形成超过几十万首曲目的原创作品库，对于音乐流行趋势的把握有着极为重要的指引作用。本项目以阿里音乐用户的历史播放数据为基础，期望可以通过对阿里音乐平台上每个阶段艺人的试听量的预测，挖掘出即将成为潮流的艺人，从而实现对一个时间段内音乐流行趋势的准确把控。

B. 项目评估

对数据进行预处理和可视化分析，设计合适的模型（如时间序列分析之 ARIMA 结合神经网络等），根据前 6 个月数据，预测后两个月的艺人每日播放量。利用 阿里官方的真实数据呈现的排名，评估预测效果。

C. 阿里音乐历史排名

5476 支队伍中，最优成绩第 1 名，目前 *34* 名（截止于 6 月 9 日）

II 中期

比赛要求是音乐播放量预测，给定的数据包括前 6 个月的用户行为表和歌曲艺人表，表的定义如下：

user_id	song_id	gmt_create	action_type	Ds
xxxxx	sbgcvd	142640600	1	20150315

表 1 用户行为表

song_id	artist_id	publish_time	song_init_plays	language	gender
Sbgcvdkj	Xaxaxaxa	20150325	0	100	1

表 2 歌曲艺人表

artist_id	Plays	Ds
Xaxaxaxa	5000	20150325

表 3 需要提交的表格包含后两个月每个艺人的播放量预测

数据处理过程中，因为不过多考虑用户行为，我们的做法就是先将给定的两个表格聚合，求出每个艺人在每一天的歌曲播放量、下载量、收藏量，然后利用这些数据作为预测的输入数据。

为了建立数据分析模型，我们对数据进行可视化分析： 比如对于歌手 03c6699ea836decbc5c8fc2dbae7bd3b 我们绘制出了该歌手每一首歌曲的播放收藏和下载数据，然后计算了他所有歌曲 183 天的播放、下载、收藏和播放的 1001og 值。如果把所有歌手的图放一块，我们可以发现，除了发布歌曲的节点外，多数变化具有周期性，而且收藏和下载比在一定程度上可以衡量一个歌手的发展潜力 现在的算法模型比较简单，用到的思想也就是简单的时间序列方法。

1)将前一段时间的数据求均值，作为后继的结果，然后移动窗口，再求均值

2)调整窗口的大小可以得到不同的结果，不断调整可以得到一个较好的结果

3)不同数据的波动性不同对结果的影响较大，我们旨在产生较为稳定的数据，所以方差小的窗口比大的好

4)下载量和收藏量数据也要利用起来

5)考虑去掉窗口中的最大值和最小值，增加结果的稳定性

算法伪代码

获得每个歌手 183 天内的所有歌曲的单天播放，收藏，下载数

对每个歌手最后多个窗口去掉最值的数据求方差

选择参数：播放下载比，为 183 天的收藏均值除以下载均值的结果

设置合适的窗口大小

while (60 天数据没有生成完):

    滑动窗口;

    去最值;

    计算均值 m;

    m = m*h(播放下载比);#h()为关于播放下载比的一个函数，还在设计之中

    设置 m 为当前播放量;

III 终期

A. 背景

阿里音乐拥有数百万的曲库资源，拥有数亿人次的用户试听、收藏等行为。在原创艺人和作品方面，拥有数万的独立音乐人，每月上传上万个原创作品，形成超过几十万首曲目的原创作品库，对于音乐流行趋势的把握有着极为重要的指引作用。本项目以阿里音乐用户的历史播放数据为基础，期望可以通过对阿里音乐平台上每个阶段艺人的试听量的预测，挖掘出即将成为潮流的艺人，从而实现对一个时间段内音乐流行趋势的准确把控。

B. 问题重述

此问题的目标是对阿里音乐的播放量进行预测，详细说来，给定多个用户长达 6 个月的播放数据，包括用户行为表和歌曲艺人表，结果是每个艺人后两个月的播放量，形式如下：

user_id	song_id	gmt_create	action_type	Ds
xxxxx	sbgcvd	142640600	1	20150315

表 4 用户行为表

song_id	artist_id	publish_time	song_init_plays	language	gender
Sbgcvdkj	Xaxaxaxa	20150325	0	100	1

表 5 歌曲艺人表

artist_id	Plays	Ds
Xaxaxaxa	5000	20150325

表 6 需要提交的表格包含后两个月每个艺人的播放量预测

C. 方法综述

1) 特征选取

这个预测问题涉及到多个属性，包括播放量、下载量、收藏量、粉丝数（听众），以及单首歌的播放量，歌手性别，歌的发行时间和语言等。特征选取是模型建立中极其重要的一环，通过可视化分析，我们发现对播放量影响直接的特征是播放量、下载量、收藏量，所以我们选取这三个特征。新歌显然会对后面的播放量产生很大影响，但是在后期预测过程中我们没有任何方法能捕捉到新歌发行事件，所以这个不做考虑。

2) 朴素的均值方法

我们首先分别求出每个歌手每一天的所有歌曲播放量、下载量和收藏量和，并做可视化操作，如下：

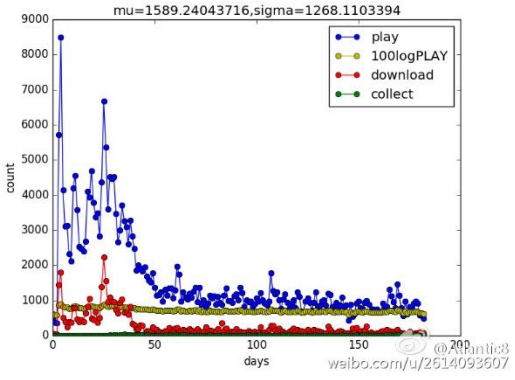


图 1

通过分析发现，整体播放量的波动影响是有少数几首歌的火爆引起的，可能是新歌也可能是老歌，由于对于新歌和老歌的流行预测是非常困难的，我们只关注预测稳定趋势，而不考虑播放量暴增或者暴减的突发事件。算法的主要

思想是使用给定数据后 n 天的播放量的均值作为后面的预测值，且后面所有的预测值都一样，另一点需要注意的是我们在计算均值之前需要将 n 天的播放量去除最值，计算公式如下：

$$x = \frac{1}{n-2}(\sum_{i=1}^{n} x_i - x_{max} - x_{min})$$

3)基于滑动窗口的均值预测方法

基于滑动窗口的均值预测方法是建立在朴素的均值基础上的方法。一般地，如果给定数据的后几天数值会出现的波动较大，波动会逐渐消失，数据变化会渐趋平缓。为了能够体现这样的变化趋势，我们利用滑动窗口的思想，公式如下：

$$x_{n+1} = \frac{1}{n-2}(\sum_{i=1}^{n} x_i - x_{max} - x_{min})$$

上述方法只用到了播放量这一种数据，而下载量和收藏量也一定与后来的播放量有关系，但是由上节图示可以看出收藏量与播放量的相关性较小，而下载量与播放量的相关性则较大。本方案中，我们加入播放下载比这样一个参数，具体算法描述如下：

```
# 获得每个歌手 183 天内的所有歌曲的单天播放，收藏，下载数
preprocess;
# 计算播放下载比
compute h;
while (not finish):
    get multi-window with different length;
    for each window:
        remove extreme value;
        calculate mean and variance;
    select window x with smallest variance;
    set m = mean of window x;
    #h()为关于播放下载比的一个函数
    m = m * h(播放下载比);
    set m as current play;
```

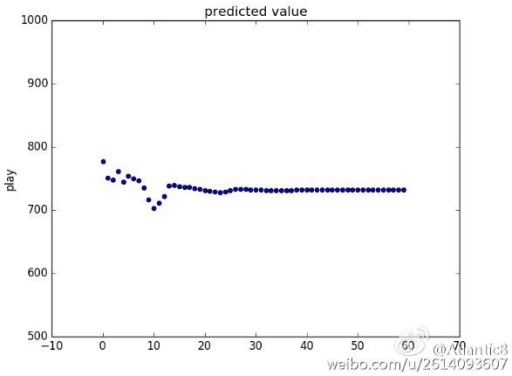
D. 实现与结果

1)实现

算法用 python 语言实现，系统环境为 windows10，主要分为预处理阶段和预测阶段。其中，预处理阶段将给定的两个原始表格数据聚合，得到每个歌手每一天的所有歌曲播放量、下载量和收藏量；预测阶段是利用预处理阶段的数据，应用相应的预测算法求出预测值。

2)结果

对算法结果可视化如下图所示，该图是对上文中出现图片对应歌手数据的预测。



第一赛季最优成绩为第 1 名，最终成绩为 103 名，共 5000+支队伍。