

CSDN

博客 (<http://feed.csdn.net/?ref=toolbar>) 学院 (<http://edu.csdn.net/?ref=toolbar>)

下载 (<http://download.csdn.net/?ref=toolbar>) 更多 ▾

登录 (<https://passport.csdn.net/account/login?ref=toolbar>) 注册 (<http://passport.csdn.net/account/mobileregister?ref=toolbar&action=mobileRegister>)

<http://whiteblog.csdn.net/postedit?ref=toolbar> /activity?utm_source=csdnblog1)

阿里音乐流行趋势预测大赛一起做-（7）初赛总结之用户分类

转载 2017年02月16日 14:55:48

标签: 预测 (<http://so.csdn.net/so/search/s.do?q=预测&t=blog>) / 音乐 (<http://so.csdn.net/so/search/s.do?q=音乐&t=blog>) / 阿里 (<http://so.csdn.net/so/search/s.do?q=阿里&t=blog>) / 趋势 (<http://so.csdn.net/so/search/s.do?q=趋势&t=blog>)

190

xiaopihaierletian (<http://blog.csdn.net/xiaopihaierletian>)

+关注

(<http://blog.csdn.net/xiaopihaierletian>)

原创 粉丝 喜欢 码云 未开通

458 110 1 (<https://github.com/xiaopihaierletian>)

他的最新文章

更多文章 (<http://blog.csdn.net/xiaopihaierletian>)

教你彻底学会动态规划——进阶篇 (<http://blog.csdn.net/xiaopihaierletian/article/details/78415577>)

教你彻底学会动态规划——入门篇 (<http://blog.csdn.net/xiaopihaierletian/article/details/78415125>)

北邮Google大牛本科研究生总结~ (<http://blog.csdn.net/xiaopihaierletian/article/details/78267438>)

音乐预测大赛的初赛今天轰轰烈烈地谢幕了，先来看看排名前10的队伍吧

第 1 赛季排行榜				
排名	参赛者	所在组织	评分	最优成绩提交日
1	<div>datahacker</div>	南京邮电大学先进技术研究院	15488	2016-06-14
2	<div>Mat_hot</div>	华南理工大学	15485	2016-06-12
3	<div>merchandising</div>	USTC、SBS Tech	15428	2016-06-09
4	<div>三千世界</div>	中国科学技术大学	15419	2016-06-14
5	<div>cike</div>	华南理工大学	15406	2016-06-13
6	<div>14jzhan</div>	汕头大学	15392	2016-06-14
7	<div>noleya_liuyang</div>	南京农大	15390	2016-06-14
8	<div>左手诗句，右手数据</div>	重庆邮电大学	15384	2016-06-13
9	<div>就是gan</div>	重庆邮电大学	15381	2016-06-13
10	<div>章鱼gege</div>	东北大学	15328	2016-06-14

我们队的成绩在切换数据后一直没有特别大的长进，虽然成绩没有达到预期，但是初赛这些天还是涌现并实现了一些不错的思路，有一些还没有来得及实现或者优化，相比去年资金预测时候的我们已经有很大进步啦（自我安慰）。看到群里不少人在询问大神团队的经验方法，虽然我们排名比较靠后，远远算不上大神，但是也有一些收获，故在此总结一下我们队用到的比较重要的特征或者方法，希望给新人有所帮助！

我们队的宗旨

首先说明一下我们团队比赛的宗旨：

- 1、尽量用模型实现，少些人工成分。因为去年比赛没有经验，所以很多时间都在人工调整参数（算是规则吧），有时候就是臆想或瞎猜，现在想想其实收获并不大，还很浪费时间，打比赛就是要学习一套思路和方法。
- 2、坚持每天提交成绩，每天有进步。每天提交成绩是我们立下的目标，比赛下来2个月每天坚持也是一种挑战，这也是一种逼自己去想去思考的办法。每次提交成绩都要有针对性，要么是尝试某个特征，要么是尝试某个模型，或者某种新思



在线课程

SDCC 2017 前端技术实践线上峰会 (<http://edu.csdn.net/course/detail/6087>)

MyOJ在美国点还漂流移动端的最佳实践 (http://edu.csdn.net/huiyiCourse/detail/594?utm_source=blog9)

C语言大型软件设计的面向对象 (http://edu.csdn.net/huiyiCourse/detail/594?utm_source=blog9)

热门文章

路，切忌盲目地手动改改就提交成绩，掩耳盗铃啊。

3、多去尝试新的思路和方法。一步步深入下去你会发现很多有趣的东西（后续博客中会具体介绍）。

好啦，废话不多说，下面进入正题。下面给出我们队一步步做用户分类（也称为用户画像）的思路和方法。

用户分类初步

资金预测的时候大熊老师做过对余额宝用户的分类，这点我在《阿里音乐流行趋势预测大赛一起做-（5）温故知新》(<http://blog.csdn.net/electech6/article/details/51512766>)里详细提到，有兴趣可以回看一下。这个思路估计大部分队伍都有，那么具体如何做呢。

1、首先大家都能想到的就是粉丝用户和非粉丝用户。我们可以简单的认为粉丝用户就是：某个用户对一首歌有下载，并且播放总量超过一定的阈值，则认为该用户是粉丝用户；否则是普通非粉丝用户。当然也可以定义为有收藏，并且播放总量超过一定的阈值，则认为是粉丝用户，总之就是收藏、下载、播放量的组合选择。

2、我们需要得到一个统计每个用户对其所听的每个歌曲的总收藏量、总下载量、总播放量的一个表（见后面user_song_index表）。如下图所示：

user_id	song_id	download_num	collect_num	play_num
0000110e00f7c80d4d9577bb41	0000110e00f7c82227abfc03b9	0	0	1
0000110e00f7c848832614076d	0000110e00f7c8638b384e3aef	0	0	1
0000110e00f7c86723ab4bd626	0000110e00f7c86c13ac393fc7	1	0	33
0000110e00f7c87ee85bd73079	0000110e00f7c88c3ceb5cb39b	1	0	7
0000110e00f7c89b28a121b0b5	0000110e00f7c8a252643c4d52	1	0	19
0000110e00f7c8c1c8390264e5	0000110e00f7cef225625ab75	0	0	1
0000110e00f7cf6396be06d56	00004931fe12df17be7a08d59c	1	0	1
00004931fe12df241b146e6b7c	00004931fe12df46e583f84731	0	0	1
00004931fe12df5e99996eda54	00004931fe12dfe453161191eb	0	0	1
00004931fe12dfce4f9f1154c		1	0	1

图中红色和绿色框内分别表示两个不同用户对其所听的所有歌曲的收藏、下载、播放总量的统计表。

3、这部分的代码如何实现呢？这是很多人最关心的问题。我们队实现的Python (<http://lib.csdn.net/base/11>)代码如下：

```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4
5 # 生成用户-歌曲-行为统计表
6
7 tianchi_user_actions = pd.read_csv('F:/AliMusicTianChi/mars_tianchi/mars_tianchi_user_actions.csv')
8 # 删除列gmt_creat
9 del tianchi_user_actions['gmt_creat']
10 del tianchi_user_actions['Ds']
11# 将action_type做哑变量处理
12temp = pd.get_dummies(tianchi_user_actions['action_type']).join(tianchi_user_actions)
13# 对操作类型对应的列名进行重命名
14temp = temp.rename(columns={'1':'play_num',2:'download_num',3:'collect_num'})
15# 删除列action_type
16del temp['action_type']
17# 将数据按照user_id,song_id两级聚合
18grouped = temp.groupby([temp['user_id'],temp['song_id']])
19# 统计播放、下载、收藏的总量
20user_song_index = grouped.agg({'play_num':'sum','download_num':'sum','collect_num':'sum'})
21# 导出为csv表格
22user_song_index.to_csv('F:/AliMusicTianChi/data/classify_users/user_song_index.csv')
```

4、得到了这样的一个表user_song_index后，我们需要把mars_tianchi_user_actions表中按照不同的用户分别取出来，一个专门存放粉丝用户的行为记录（表fan_table），另一个专门存放非粉丝用户的行为记录（表non_fan_table）。有了表user_song_index，我们可以用如下代码实现不同用户行为记录的提取：

深度学习与计算机视觉系列(5)_反向传播与它的直观理解 (<http://blog.csdn.net/xiaopihaierletian/article/details/5522985>)

7161

tensorflow学习笔记十四：TF官方教程学习 tf.contrib.learn Quickstart (<http://blog.csdn.net/xiaopihaierletian/article/details/5522985>)

5814

Windows 下面安装和使用Python, IPython Notebook (详细步骤) (<http://blog.csdn.net/xiaopihaierletian/article/details/5522985>)

5734

新浪微博开发者平台应用申请及配置说明 (<http://blog.csdn.net/xiaopihaierletian/article/details/5522985>)

4333

tensorflow学习笔记十四：tensorflow中的tf.app.run() (<http://blog.csdn.net/xiaopihaierletian/article/details/5522985>)

3933

```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 # 用户分类：认为有下载、总播放量大于5的用户是粉丝用户，其他为非粉丝用户
5 user_song_index = pd.read_csv('F:/AliMusicTianChi/data/classify_users/user_song_index.csv')
6 # 设置用户分类规则
7 index = np.logical_and(user_song_index['download_num']>0,user_song_index['play_num']>5)
8 fan_list_tmp= user_song_index[index]
9 fan_list = sorted(set(fan_list_tmp['user_id']))
10
11 # in1d(x,y)表示得到一个x是否包含于y的布尔型数组
12 fan_index = np.in1d(tianchi_user_actions['user_id'],fan_list)
13 non_fan_index = np.logical_not(fan_index)
14 fan_table = tianchi_user_actions[fan_index]
15 non_fan_table = tianchi_user_actions[non_fan_index]
16
17 fan_table.to_csv('F:/AliMusicTianChi/data/classify_users/fan_table.csv')
18 non_fan_table.to_csv('F:/AliMusicTianChi/data/classify_users/non_fan_table.csv')
```

5、然后我们用这两个表分别运行自己的模型，然后叠加到一起就是最后的结果。

用户分类进阶

前面我们已经可以实现根据用户不同行为组合进行简单的用户分类。那么问题来了，这样分类真的好吗？什么样的分类才是好的分类？

我觉得好的分类要满足几个条件：

1、首先分类具有可解释性。针对音乐预测我们可以考虑一下针对用户这一个维度分类是否合理。我们可以考虑大部分人的听歌习惯：

- 1）下载了音乐软件后，首先收藏自己喜欢的歌曲或者艺人，下载自己喜欢的歌曲。
- 2）想听歌了，打开收藏的列表或者下载的列表的歌曲播放。
- 3）有时候心血来潮可能会试听软件推荐的歌单。
- 4）听了电视剧或电影中的主题曲、插曲觉得不错会试听。
- 5）得知某个喜欢的歌手发行了新歌会去下载、播放、收藏。
- 6）某首歌曲最近比较火，会播放听一听，如果喜欢可能会下载或收藏。

相信大部分的人习惯都是这样，所以每个用户的每天可能有稳定的播放歌曲。换句话说，就是每个用户可能下载了很多歌曲，但是经常播放的只是其中一部分。所以“用户分类初步”里使用用户一个维度进行粉丝划分并不是非常符合实际情况。

综上，建议使用user_id & song_id两个维度作为粉丝分类更为合适。

2、用户分类后的统计数据应该在同一个数量级上。举个极端点的例子，比如你按照某种方式分类后粉丝的播放量一般为每天10次左右，非粉丝的一般为每天1000次左右，由于数量级相差较大，那么两者相加时粉丝的分类结果就淹没在非粉丝的茫茫大海之中。即使你粉丝部分预测的再好也无济于事。

3、用户分类后的数据应该更容易预测。用户分类的目标就是把一个复杂问题分解成多个简单问题的叠加。所以衡量一个分类是否好坏的重要标志就是分类后的数据应该变地更平稳了、更有规律了、或者有很好的线性等，这些特点可以提高预测的精度。

后记

每个人根据不同的方法进行用户分类都会有不同的结果，有的人可能分离出容易预测的线性数据，有的人可能分离出明显的周期性数据。一千个人眼中有一千个哈姆雷特，所以如果你深入一步步地挖掘，一定会发现适合自己的分类方法。

欢迎参赛同学分享观点，祝大家在复赛中取得佳绩！

相关文章推荐

阿里音乐流行趋势预测大赛一起做- (7) 初赛总结之用户分类 (<http://blog.csdn.net/electech6/article/details/52599>)

音乐预测大赛的初赛今天轰轰烈烈地谢幕了，先来看看排名前10的队伍吧我们队的成绩在切换数据后一直没有特别大的长进，虽然成绩没有达到预期，但是初赛这些天还是涌现并实现了一些不错的思路，有一些还没有...

 electech6 (<http://blog.csdn.net/electech6>) 2016年06月14日 13:12 12599

阿里音乐流行趋势预测大赛一起做- (3) 思路 (<http://blog.csdn.net/electech6/article/details/5499>)

最近几个月忙着写论文，比赛给抛到脑后了（汗颜。。）。昨天阿里给了短信通知，说今天可以提交成绩了，于是抓紧搞起，结果也没赶上截止时间（早上10点）提交。无论如何今天也要提交一次成绩试试水。比赛的详细思路...

 electech6 (<http://blog.csdn.net/electech6>) 2016年05月17日 22:32 5499



11.11 程序员该买些什么犒劳自己？

这是一份11.11程序员版爆款剁手清单！万万没想到，他们买了这些...

(http://www.baidu.com/cb.php?c=lgF_pyfqHmknjDLnJT0IZ0qnfK9ujYzP1nsrjDz0Aw-5Hc3rHnYnHb0TAq15HfLPWRznjb0T1dhmW9WPY7WPYnLrHfvrjck0AwY5HDdnH0vnWRzrHD0lgF_5y9YIZ0IQzq-uZR8mLPbUB48ugfEXyN9T-KzUvdEIA-EUBqbugw9pysEn1qdlAdxTvqdThP-5yF_UvTkn0KzujY4rHb0mhYqn0KsTWYS0ZNGujYkPHTYn1mk0AqGujYknWb3rjDY0APGujYLnWm4n1c0ULI85H00TZbqnW0v0APzm1YdrjRdP6)

阿里音乐流行趋势预测大赛一起做- (2) weka初识 (<http://blog.csdn.net/xiaopihaierletian/article/details/5126>)

本次比赛第一赛季我们团队大赛先用weka对数据进行预处理。weka是一款开源的，免费的，基于JAVA环境下开源的机器学习（machine learning）以及数据挖掘（data mining）软...

 xiaopihaierletian (<http://blog.csdn.net/xiaopihaierletian>) 2017年02月16日 14:51 126

阿里音乐流行趋势预测大赛一起做- (1) 介绍 (<http://blog.csdn.net/electech6/article/details/6005>)

动机去年在同学推荐下参加了阿里的“资金流入流出预测大赛”，当时对数据挖掘的了解仅存在模糊的概念上，看了那本《大数据时代》后热血沸腾了几天，趁着这三分钟热度还没冷，就在什么都不懂的情况下报名开始搞了。我...

 electech6 (<http://blog.csdn.net/electech6>) 2016年04月09日 11:11 6005

阿里音乐流行趋势预测大赛一起做- (1) 介绍 (<http://blog.csdn.net/xiaopihaierletian/article/details/300>)

动机去年在同学推荐下参加了阿里的“资金流入流出预测大赛”，当时对数据挖掘的了解仅存在模糊的概念上，看了那本《大数据时代》后热血沸腾了几天，趁着这三分钟热度还没冷，就在什么都不懂的情况下报名开始...

 xiaopihaierletian (<http://blog.csdn.net/xiaopihaierletian>) 2017年02月16日 14:50 300



人人都能看懂的 AI 入门课

本课程将讲述人工智能的现状、应用场景和入门方法，并通过运用 TensorFlow，使得受众能清晰了解人工智能的运作方式。

(http://www.baidu.com/cb.php?c=lgF_pyfqHmknjcvPjm0IZ0qnfK9ujYzP1ndnHfz0Aw-5Hc4nj6vPjm0TAq15Hf4rjn1n1b0T1YkmvnsPhmYmH-)

hujRzuAn40AwY5HDdnH0vnWRzrHc0lgF_5y9YIZ0lQzqMpgwBUvqoQhP8QvIGIAPCmgfEmvq_lyd8Q1R4uWI-n16kPWKWrHnvnHRvvnNBuyD4PHqdlAdxTvqdThP-5HDknHK4mhkEusKzujY4rHb0mhYqn0KsTWYs0ZNGUjYkPHTYn1mk0AqGujYkn10snjf10APGujYLnWm4n1c0ULI85H00TZbqnW0v0APzm1Ydrj6kPs)

阿里音乐流行趋势预测大赛一起做-（3）思路 (http://blog.csdn.net/xiaopihaierletian/art...

最近几个月忙着写论文，比赛给抛到脑后了（汗颜。。）。昨天阿里给了短信通知，说今天可以提交成绩了，于是抓紧搞起，结果也没赶上截止时间（早上10点）提交。无论如何今天也要提交一次成绩试试水。比赛的详细思路...

 xiaopihaierletian (http://blog.csdn.net/xiaopihaierletian) 2017年02月16日 14:52 222

阿里音乐流行趋势预测大赛一起做-（5）温故知新 (http://blog.csdn.net/electech6/articl...

比赛进行到现在已经有十天了，我们队伍已经进入瓶颈期了，具体表现就是： 1、每次线下预测结果都感觉不错，应该可以提升成绩吧，至少和现在成绩相当吧，结果10点成绩出来就傻眼了，成绩一直掉掉掉。 2、黔...

 electech6 (http://blog.csdn.net/electech6) 2016年05月27日 11:17 4114

阿里音乐流行趋势预测大赛一起做-（4）成绩提交 (http://blog.csdn.net/xiaopihaierletia...

按照我上一篇博客中所描述的思路，我们仅用artist_id和艺人当天的播放量Plays，以及日期Ds三个维度的数据作为训练集合，对每个艺人分别使用时间序列算法进行预测，提交了成绩。本来没报多大希望，只...

 xiaopihaierletian (http://blog.csdn.net/xiaopihaierletian) 2017年02月16日 14:52 286


阿里音乐流行趋势预测大赛一起做-（6）小结 (http://blog.csdn.net/electech6/article/det...

最近一直在忙毕业的事情，比赛也落下不少。今天第一次切换了数据，在此对之前的工作做个简单的总结。感谢组委会岱月邀请我写自己的参赛经历，《天池小白成长记》在阿里的“天池大数据科研平台”上发布了，还是蛮激动...

 electech6 (http://blog.csdn.net/electech6) 2016年06月07日 14:59 3126


阿里音乐流行趋势预测大赛一起做-（5）温故知新 (http://blog.csdn.net/xiaopihaierletia...

比赛进行到现在已经有十天了，我们队伍已经进入瓶颈期了，具体表现就是： 1、每次线下预测结果都感觉不错，应该可以提升成绩吧，至少和现在成绩相当吧，结果10点成绩出来就傻眼了，成绩一直掉掉掉。 2...

 xiaopihaierletian (http://blog.csdn.net/xiaopihaierletian) 2017年02月16日 14:53 229


阿里音乐流行趋势预测大赛一起做-（4）成绩提交 (/electech6/article/details/51459292)

按照我上一篇博客中所描述的思路，我们仅用artist_id和艺人当天的播放量Plays，以及日期Ds三个维度的数据作为训练集合，对每个艺人分别使用时间序列算法进行预测，提交了成绩。本来没报多大希望，只...

 electech6 (http://blog.csdn.net/electech6) 2016-05-20 10:47 4342

阿里音乐流行趋势预测大赛-赛后思路整理 (/q383700092/article/details/51986280)



赛题介绍阿里音乐用户的历史播放数据预测阿里音乐平台艺人的未来60天的播放量最后top15 1从图入手：查看规律与趋势 2提出问题：为什么会出现特别高的播放量？是否有一定的节假日工作日规律...

 q383700092 (http://blog.csdn.net/q383700092) 2016-07-21 20:15 1447

阿里音乐流行趋势预测大赛—浅尝辄止（三） (/qq_14959801/article/details/52232204)



这篇博文是在阿里音乐大赛（一）和阿里音乐大赛（二）之后的，参考到以为博主的文章和代码给我带来了启发，

发现了很多Pandas的新用法，所以在这里做一个记录总结；关于阿里音乐大赛的相关方法介绍可以参考...

 qq_14959801 (http://blog.csdn.net/qq_14959801) 2016-08-17 16:34  309



阿里音乐流行趋势预测大赛—浅尝辄止（二） ([/qq_14959801/article/details/52054948](http://blog.csdn.net/qq_14959801/article/details/52054948))

本篇博文接上一篇博文浅尝辄止（一）的内容，这里主要介绍竞赛给的数据中时间戳的处理方法，时间戳是形如“1426406400”形式的一组时间计数，我们需要将其转化为正常的时分秒的形式，然后再将转化后的结果...

 qq_14959801 (http://blog.csdn.net/qq_14959801) 2016-07-28 14:23  341


阿里音乐流行趋势预测大赛—浅尝辄止（一） ([/qq_14959801/article/details/52054944](http://blog.csdn.net/qq_14959801/article/details/52054944))

如火如荼的阿里音乐流行趋势预测大赛终于落下帷幕，作者未能全力以赴的好好参与其中拿个名次实在遗憾，作为菜鸟新人还是有不少心里上过不去的坎的，希望接下来能结识更多的爱好者，共同学习进步，争取在下次比赛的时候...

 qq_14959801 (http://blog.csdn.net/qq_14959801) 2016-07-28 14:23  403



天池_阿里音乐流行趋势预测大赛（1）——赛题分析 ([/lipengcn/article/details/51553468](http://blog.csdn.net/lipengcn/article/details/51553468))

本文以天池大数据竞赛的阿里音乐流行趋势预测大赛为背景，将机器学习实战的背景、模型、算法、代码和结果等都整理下来，放在博客中，算是对自己知识的整理吧，有兴趣的朋友也可以看看一起讨论学习。由于很多比赛和项...

 lipengcn (<http://blog.csdn.net/lipengcn>) 2016-06-01 21:23  1782



阿里音乐流行趋势预测大赛 # 第一赛季第21名解决方案 ([/zxjhdn/article/details/51957820](http://blog.csdn.net/zxjhdn/article/details/51957820))

强调内容# 欢迎使用Markdown编辑器写博客本Markdown编辑器使用StackEdit修改而来，用它写博客，将会带来全新的体验哦： Markdown和扩展Markdown简洁的语法代码块高亮...

 zxjhdn (<http://blog.csdn.net/zxjhdn>) 2016-07-19 17:28  2256



CodeM美团点评编程大赛初赛A轮 C.倒水【二分+分类讨论】 ([/mengxiang000000/articl...](http://blog.csdn.net/mengxiang000000/article/details/51957820))

[编程题] 倒水时间限制：1秒空间限制：32768K 有一个大水缸，里面水的温度为T单位，体积为C升。另有n杯水（假设每个杯子的容量是无限的），每杯水的温度为t[i]...

 mengxiang000000 (<http://blog.csdn.net/mengxiang000000>) 2017-06-19 20:08  265



“第一届腾讯高校算法大赛”初赛失败总结 ([/earl211/article/details/72900406](http://blog.csdn.net/Earl211/article/details/72900406))

经过一个月的无用功，竟然都没有走进复赛，失望的心情难以平复。在知道这个消息的时候立即写一篇失败总结，让自己获取一些经验。这篇总结我将按照《精进》这本书里面提到的几个方面进行总结：信息，预期，结果，进度...

 Earl211 (<http://blog.csdn.net/Earl211>) 2017-06-07 15:37  1386

Anyoffice -HTML5大赛 悦心（基于H5开发安卓音乐app）-项目总结 ([/u014345282/articl...](http://blog.csdn.net/u014345282/article/details/51957820))

项目在线演示地址：<http://rose111.applinzi.com/> 悦心 悦心,一款音乐播放器应用，由“女立方”团队开发。目前，较为流行的音乐播放器有QQ音乐、网易云音乐、多米音乐等。“悦...

 u014345282 (<http://blog.csdn.net/u014345282>) 2016-04-08 16:49  1612