CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 40 卷第 9 期 2018 年 9 月
Vol. 40, No. 9, Sep. 2018

文章编号:1007-130X(2018)09-1703-07

基于时间序列的音乐流行趋势预测研究*

郁伟生[1],邓 伟[1],张 瑶[2],李蜀瑜[1,2]

(1.陕西师范大学网络信息中心,陕西 西安 710119;2.陕西师范大学计算机科学学院,陕西 西安 710119)

摘 要:在大数据环境下,对音乐及听众的历史数据进行分析,可以实现对音乐流行趋势较为准确的预测。在 STL、Holt-Winters 分解模型的基础上,提出 TSMP 算法。该算法从长期趋势和周期两方面进行分析,对长期趋势编码和分类并基于类别最优值选择法对音乐流行趋势进行预测。基于 TSMP 算法,进而提出 E-TSMP 算法,该算法基于子序列模式匹配法及对近期发布新专辑的附加处理,实现更精准的预测。在清华大学和阿里云天池大数据竞赛平台承办的"2016 中国高校计算机大赛——大数据挑战赛之阿里音乐流行趋势预测"比赛中,参赛团队凭借提出的 E-TSMP 算法对 2016 年 9 月~10 月艺人的播放量实现了较好的预测,并在此次比赛中夺得亚军。

关键词:时间序列;音乐流行趋势;类别最优值选择;子序列模式匹配

中图分类号:TP301 **文献标志码**:A

doi:10.3969/j.issn.1007-130X.2018.09.024

Music popular trends prediction based on time series

YU Wei-sheng[1], DENG Wei[1], ZHANG Yao[2], LI Shu-yu[1,2]

(1. Network Information Center, Shaanxi Normal University, Xi'an 710119;
2. School of Computer Science, Shaanxi Normal University, Xi'an 710119, China)

Abstract: In big data environment, analyzing the historical data of music and audiences can achieve accurate prediction of music popular trends. Based on the STL and Holt-Winters decomposition models, a Time Series based Music Prediction (TSMP) algorithm is proposed. The TSMP algorithm analyzes data from both long-term trends and periods, codes and categorizes the long-term trends, and uses the category optimal value selection method to predict the music popular trends. Based on the TSMP algorithm, the Extend TSMP (E-TSMP) algorithm is proposed, which is based on the subsequence pattern matching method and the additional processing of the newly released new album to achieve more accurate prediction. In the "2016 Chinese University Computer Contest-Big Data Challenge Ali music popular trends prediction competition" hosted by Tsinghua University and Tianchi big data competition platform of Ali cloud, the participating team uses the proposed E-TSMP algorithm to achieve good prediction for artist's play times from september to october in 2016 and won the second place in this competition.

Key words: time series; music popular trends; category optimal value selection; sub-sequence pattern matching

1 引言

音乐是反映人类现实生活情感的一种艺术,当下音乐的流行趋势成为了众人关注的话题。在如今的大数据时代,音乐听众将会决定音乐的流行趋势。听众在众多音乐平台上试听、下载、收藏、分享音乐,以及在各大社交网络、视频网站、贴吧、论坛

* 收稿日期:2017-06-14;修回日期:2017-08-15
基金项目:国家自然科学基金(41271387);中央高校基本科研业务费项目(GK201703055)
通信地址:710119 陕西省西安市西长安街 620 号陕西师范大学长安校区网络信息中心
Address: Network Information Center, Chang'an Campus, Shaanxi Normal University, 620 West Chang'an St, Xi'an 710119, Shaanxi, P. R. China

上对音乐进行关注、评论、转发、点赞,这些行为反映出听众对音乐的喜好。当今音乐的流行趋势,可以通过对听众的喜好趋向进行深度挖掘和分析预测获得。利用庞大的曲库资源和用户行为形成音乐大数据,通过精准的大数据分析,可以有效预测音乐的潮流走向,真正实现听众喜好的聚合决定音乐的流行发展趋势。

阿里音乐经过7年发展,现已拥有了数百万的曲库资源,及数亿次的用户试听、下载、收藏等行为。阿里举办的音乐流行趋势预测大赛基于阿里云平台强大的数据计算能力,通过用户的历史行为数据,预测下一阶段艺人播放量,挖掘出即将成为关注热点的艺人,从而准确把控未来音乐的流行趋势。本文以大赛复赛提供的音乐用户从2016年3月~8月的历史播放数据为基础,借鉴STL、Holt-Winters模型中的分解思想,分别从长期趋势、周期两个方面进行分析,对长期趋势编码和分类,并基于类别最优值选择和子序列模式匹配法,提出了E-TSMP(Extend-Time Series based Music Prediction)算法,最终在阿里音乐平台上实现了对2016年9月~10月艺人播放量较为准确的预测。

2 准备工作

对音乐流行趋势的预测可以采用时间序列、回归等预测模型来实现[1-7]。Li[8]提出的自回归积分滑动平均模型ARIMA(Auto Regressive Integrated Moving Average model),虽能很好地根据动态数据及自身的相关特征进行预测,但因ARIMA中差分次数d和p、q参数的选取不具备通用性,需对每个艺人数据预处理分类后逐一进行参数调整。Chatfield等[9]提出的三次指数平滑TOES(Three Order Exponential Smoothing)模型可以对同时含有趋势和季节性的时间序列进行预测,但该模型对数据集和时间段的选取比较敏感,此外针对高于二阶拟合的曲线会出现不可控的发散状态。Cleveland等[10]提出的STL分解(Seasonal and Trend decomposition using Loess)模型虽兼具通用性和鲁棒性,但只适合加法模型,且不能根据数据突然变化进行自动处理。Jain等[11]提出的递归神经网络RNN(Recurrent Neural Network)虽可以根据之前的数据给出相应反馈,并采用非线性动力系统,但收敛性差,即使加入相应的特征,预测效果也不理想。

采用ARIMA等时间序列模型中标准的时间序列函数对每个艺人的日播放量曲线进行拟合和预测,若需要预测的日期较长,容易出现过拟合等问题。在大数据环境下,对音乐进行流行趋势预测,应当对模型建立、算法设计、实验优化等方面进行综合考虑。

预测效果表示预测值和真实值之间的差距,通常定义一个打分函数来衡量预测效果。因此,阿里大赛给出相应的评估指标,用于判断预测值的准确度。设艺人j在第h天的实际播放量为$D_{j,h}$,艺人集合为A,算法预测得到艺人j在第h天的播放量为$S_{j,h}$,则艺人j的归一化方差σ_j可以根据艺人N天的实际播放量和预测值的方差求得:

$$\sigma_j=\sqrt{\frac{1}{N}\sum_{k=1}^{N}((S_{j,h}-D_{j,h})/D_{j,h})^2}$$

归一化方差σ_j反映了预测结果S_j和实际播放量D_j之间的差距,$(1-\sigma_j)$值越大,表示预测越精准。根据当前艺人j的每日实际播放量相加后的算术平方根求对应权重φ_j:

$$\varphi_j=\sqrt{\sum_{h=1}^{N}D_{j,h}}$$

定义打分函数F:

$$F=\sum_{j\in A}(1-\sigma_j)*\varphi_j$$

并以此作为预测标准,F值越大,预测值越接近实际值。

3 音乐流行趋势预测算法

在音乐的流行趋势预测过程中,借鉴了STL、Holt-Winters模型中的分解思想,在类别最优值选择法的基础之上,提出TSMP(Time Series based Music Prediction)算法。为了实现对音乐流行趋势更为精准的预测,基于子序列模式匹配法及对近期发布新专辑的附加处理,提出E-TSMP算法,针对2016年8月中下旬某些艺人的日播放量突然成倍增加的情况,有效地解决了该类情况下预测趋势后续走向问题。

3.1 TSMP算法

为了实现音乐更精准的流行趋势预测,需要对每个时间序列分别从长期趋势、周期、随机干扰项三部分进行分析。因为随机干扰项的不确定性,因此所有的研究都忽略了该部分。

阿里比赛提供了用户行为信息和歌曲艺人信息,其中用户行为信息记录用户对歌曲进行的播放、下载、收藏等操作,歌曲艺人信息包含歌曲专辑

收录时间、初始播放量、歌曲语言、歌曲类别等内容。只有在提供的众多信息中提取出有效信息,才能更好地对数据进行分析,实现对每个艺人音乐播放量更为精准的预测。因此,把信息处理成每个艺人对应的日播放量序列、周播放量均值序列、月播放量均值序列及日变化率序列,对数据进行的预处理为 TSMP 算法的实现奠定了基础。

在对每个艺人日播放量曲线进行拟合和预测的过程中,曲线趋势发展成为了至关重要的问题,而重要的外部事件是影响曲线趋势发展的一大关键因素,比如:预测期间艺人发布新专辑、开演唱会、参加选秀节目等。在排除外部事件干扰的前提下,计算每个艺人播放量的月平均值、周平均值、日平均值并进行编码处理,这些编码处理的均值可以作为每个艺人播放量趋势的预测。定义基本趋势和增量趋势作为编码规则,例如月度编码:若当月播放量均值高于前一月均值,则基本趋势对应的编码值为 1,否则为 0;若当月播放量均值高于前一月均值,则以当月均值除以上月均值的商值取整作为增量趋势编码值,否则增量趋势编码值为上月均值除以当月均值的商值取整。具体编码过程如图 1 所示。

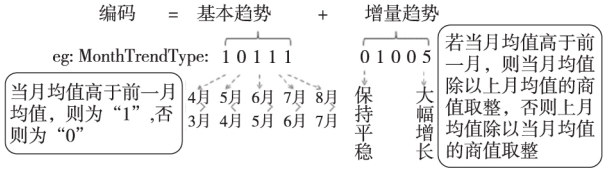


Figure 1 Coding method

图 1 编码方法

根据月编码、周编码、日编码中的基本趋势和增量趋势两部分,基于 k-means 聚类算法[12,13],最终将阿里提供的数据集划分成 24 个基本类别。通过对不同类别进行大量测试与分析,最终采用类别最优值选择法对不同类别的艺人进行日播放量预测。类别最优值选择法的思想是:选取时间序列的某个特征值作为其预测值[14],例如百分位数、后 3 天均值、后 7 日均值等,构成的如图 2 所示的候选方法预测集合用来选取特征值。

大多数艺人的时间序列相对比较平稳,实验证明,若预测的天数较长,与波动性的曲线相比,均值的预测效果更好。采用 3～7 月数据作为训练集,8 月数据作为测试集,进行预测值的特征值选取,将打分函数 F 计算的结果作为依据,并进行最大化选取,以便做最优预测。

对任意类 C_k,候选方法集为 $S=\{s^{(1)},s^{(2)},\cdots,s^{(n)}\}$,最优预测方法的选取公式为:

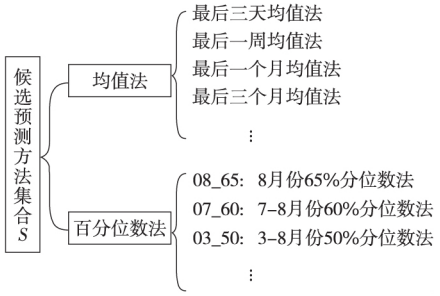


Figure 2 Candidate prediction method set based on eigenvalue

图 2 基于特征值的候选预测方法集合

$$s=\arg\max_{s^{(i)}\in S}\left[\left(1-\sqrt{\frac{1}{N}\sum_{h=1}^{N}\left(\frac{s_{j,h}^{(i)}-D_{j,h}}{D_{j,h}}\right)^2}\right)*\sqrt{\sum_{h=1}^{N}D_{j,h}}\right]$$

其中,$s_{j,h}^{(i)}$ 为采用方法 $s^{(i)}$ 预测艺人 j 在第 h 天的日播放量,$D_{j,h}$ 为艺人 j 在第 h 天的实际日播放量。针对某个类别 C_k 最优值选择算法 COVSA (Category Optimal Value Selection Algorithm)的伪代码如下所示:

算法 1 类别最优值选择算法 $COVSA(S,D_{k1},D_{k2})$

输入:S:候选方法预测集合;D_{k1}:类别 C_k 内所有艺人 3～7 月内日播放量序列(训练集);D_{k2}:类别 C_k 内所有艺人 8 月内日播放量序列(验证集)。

输出:m:类别 C_k 的最优预测方法。

```
Begin
  f←0;m←s^(1) ;/*定义打分函数初始值为0,m为最优预测方法*/
  For s^(i)∈S do//遍历候选方法预测集合
    S←S\{s^(i)};
    P←s^(i)(D_k1);/*对训练集采用候选方法s^(i)得到8月份类别C_k内所有艺人的预测集*/
    f_i←F(P,D_k2);/*计算采用候选方法s^(i)后的分数f_i*/
    if f_i > f then/*分数f_i比之前方法获得的分数f高,重设s^(i)作为最优预测方法*/
      f←f_i;
      m←s^(i);
    end
  end
  Return m;//输出最优预测方法。
End
```

随机选出类别 C_k 中 3～8 月艺人 j 的音乐播放量作为数据集,应用类别最优值选择算法最终输出类别 C_k 的最优预测方法 m,对于类别 k 中的其他艺人,用其最优预测方法 m 预测 9～10 月份的日播放量。具体分类规则及对应最优预测方法如图 3 所示。基于类别最优值选择方法,提出了

月基本趋势	月增量趋势	周基本趋势	周增量趋势	日基本趋势	最优方法
XXX11	XXXX0				08_65_percent
XXX11	XXXX[1-9]				08_65_percent
XXX11	XXXX[1-9]	XXX00			Last3days
XXX11	XXXX[1-9]	XXXX0	XXX[1-9]X		Last3days
XXX10	XXXX0	XXX10			Last2Weeks
XXX10	XXXX[1-9]				Last3days
XXX10	XXXX[1-9]	XXX00			Last3days
XXX10	XXXX[1-9]	XXX00	XXXX[1-9]		Last3days
	XXXX[5-9]				Last2Month*1.1
...					
...					
...					

Figure 3 Classification rules

图 3 分类规则

TSMP 算法，实现了对 9～10 月艺人总播放量的预测，伪代码如下所示：

算法 2 音乐流行趋势预测算法 $TSMP(U,A,S)$

输入：U：3～8 月用户行为数据集合；A：艺人基本信息集合；S：候选预测方法集合。

输出：P：9～10 月所有艺人总播放量预测值。

```
Begin
  P ← 0 ;//设初始预测值为 0
  (D,W,M) ← Pre(U,A) ;/* 把数据集 U,A 预处理为日、周、月播放量均值序列集 D,W,M */
  (DT,WT,MT) ← Cod(D,W,M) ;/* 对数据集 D,W,M 进行日、周、月编码形成编码序列集 DT,WT,MT */
  C_k ← Sort(DT,WT,MT) ;/* 采用 k-means 算法对 DT,WT,MT 进行划分构成类别 C_k，及与类别对应的日播放量序列集 D_{C_k} */
  For c_k ∈ C_k do//遍历分类集合
    C_k ← C_k\{c_k} ;
    P_{c_k} ← 0 ;/* 设类别 c_k 中所有艺人 9～10 月总播放量初始值为 0 */
    m ← COVSA(S,D_{c_k}) ;/* 根据类 c_k 内所有艺人日播放量序列 D_{c_k} 中 3～8 月的日播放量，由 COVSA 获得类别 c_k 最优预测方法 m */
    P_{c_k} ← m(D_{c_k}) ;/* 预测类别 c_k 中所有艺人 9～10 月总播放量 P_{c_k} */
    P ← P + P_{c_k} ;
  end
  Return P ;/* 输出 9～10 月所有艺人总播放量预测值 */
End
```

3.2 E-TSMP 算法

在阿里提供的 3～8 月份数据集里，通过分析周编码、日编码的增量趋势部分发现，在 8 月中下旬部分艺人的播放量突然成倍增加，若对此类艺人继续使用最优预测方法，最终预测结果与真实值偏差会较大。因此，借鉴其他艺人历史数据中出现过的类似曲线，提出了子序列模式匹配法 SSPMM (Sub-Sequence Pattern Matching Method)，对该类别艺人的预测方法做了相应的改进。子序列模式匹配法思想是：首先计算所有艺人 3～8 月份歌曲播放量日变化率序列，根据观察及大量实验进行验证，最终选择截取待预测艺人最后 15 天的日变化率序列并作为待匹配子序列集 ，然后根据序列集中艺人的日变化率，找出与待匹配子序列集欧氏距离最小的 5 个子序列，选取其中后续子序列变化比较平稳的 3 个子序列作为最佳子序列，最后求取最佳子序列的日变化率均值，作为待预测艺人的日变化率序列，计算出待预测艺人 9～10 月份的日播放量趋势中回落后的平稳部分的预测值。在回落过程，采用 n 段梯度法求取回落过程中的预测值，第 i 个回落点的预测值为：

$$Pre(i)=\frac{\lceil (n-1)*d_1-d_2\rceil *i}{n}$$

其中，d_1 为回落点艺人的日播放量，d_2 为预测艺人平稳点的预测值。子序列模式匹配法的具体实现过程如图 4 所示。

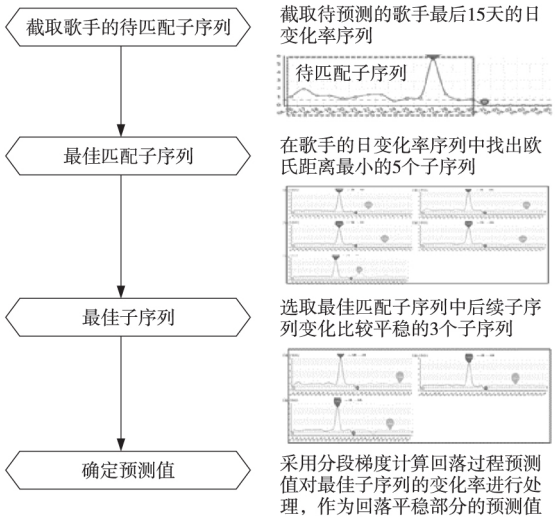


Figure 4 Sub-sequence pattern matching method

图 4 子序列模式匹配法

此外对部分类别的预测值做了附加处理 AD (Additional Processing):针对近两个月有新专辑的艺人,根据专辑的发布时间对时间做分段处理,依据以往发布专辑的艺人的播放量变化率,分别求取对应的权重因子 α,对发布专辑后一段时间的日播放量的预测值的精确度做了相应提升,AD 算法具体处理过程如图 5 所示。

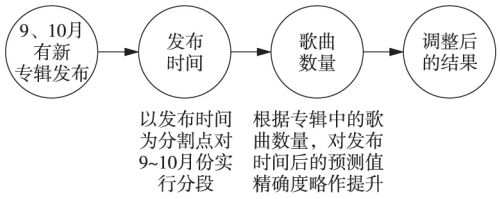


Figure 5 Additional processing

图 5 附加处理

在对所有艺人播放量的时间序列进行周期性叠加的实验中,预测效果并不理想。为了能得到良好的实验效果,在预测所有艺人的播放量的实验过程中,没有考虑周期性的叠加。基于类别最优值选择和子序列模式匹配法提出的 E-TSMP 算法,实现了对 2016 年 9～10 月艺人播放量更为精准的预测,伪代码如下所示:

算法 3 扩展音乐流行趋势预测算法 *E-TSMP*(U,A,S)

输入:U:3～8 月用户行为数据集合;A:艺人基本信息集合;S:候选预测方法集合。

输出:P:9～10 月所有艺人总播放量预测值。

```
Begin
  P ← 0 ;//设初始预测值为 0
  (D,W,M) ← Pre(U,A) ;/* 把数据集 U,A 预处理为日、周、月播放量均值序列集 D,W,M */
  (DT,WT,MT) ← Cod(D,W,M) ;/* 对数据集 D,W,M 进行日、周、月编码形成编码序列集 DT,WT,MT */
  C_k ← Sort(DT,WT,MT) ;/* 采用 k-means 算法对 DT,WT,MT 进行划分构成类别 C_k,及与类别对应的日播放量序列集 D_{C_k} */
  For c_k ∈ C_k do//遍历分类集合
    C_k ← C_k\{c_k} ;
    P_{c_k} ← 0 ;/* 设类别 c_k 中所有艺人 9～10 月总播放量初始值为 0 */
    m ← COVSA(S,D_{c_k}) ;/* 根据类 c_k 内所有艺人日播放量序列 D_{c_k} 中 3～8 月的日播放量,由 COVSA 获得类别 c_k 最优预测方法 m */
    For c_{k_j} ∈ c_k do//遍历 c_k 类所有艺人
      c_k ← c_k\{c_{k_j}} ;
      P_{c_{k_j}} ← 0 ;/* 设置 c_k 类中艺人 j 在 9～10 月总播放量的初始预测值为 0 */
      if suddenly increase then/* 艺人 j 在 8 月中下旬的播放量突然成倍增加 */
        P_{c_{k_j}} ← SSPMM(D_{c_{k_j}}) ;/* 用子序列模式匹配法预测 c_k 类中 9～10 月艺人 j 的总播放量 */
      else
        P_{c_{k_j}} ← m(D_{c_{k_j}}) ;/* 用 m 预测 c_k 类中艺人 j 在 9～10 月的总播放量 */
      end
      if publish then/* 若艺人 j 近两个月发布新专辑,略提升发布专辑后的预测值 */
        P_{c_{k_j}} ← AD(P_{c_{k_j}}) ;/* 采用 AD 算法,重新预测 c_k 类中艺人 j 在 9～10 月的总播放量 */
      end
      P_{c_k} ← P_{c_k} + P_{c_{k_j}} ;/* 预测 c_k 类中 9～10 月艺人 j 的总播放量 */
    end
    P ← P + P_{c_k} ;
  end
  Return P ;/* 输出 9～10 月所有艺人总播放量预测值 */
End
```

4 实验分析

本文以阿里比赛复赛提供的 1 000 位艺人 10 842首歌曲,从 2016 年 3 月～8 月的历史播放量作为实验数据,采用 E-TSMP 算法预测 2016 年 9 月～10 月这 1 000 位艺人歌曲总播放量。基于阿里云的 Map-Reduce 框架,使用 Java 语言,实现 E-TSMP 算法。另外,使用 Echarts 实现报表的可视化,形象地展示艺人日播放量序列变化及预测值。在复赛中,参赛团队凭借提出的 E-TSMP 算法获得大赛第 2 名,大赛复赛排行榜如图 6 所示。

在 E-TSMP 算法中的类别最优值选择过程中,将均值及百分位数作为其预测方法集,而不是将波动曲线作为预测方法集,实验表明本文方法对均值预测的效果更理想。如图 7 分别采用后 7 天均值、后 3 天均值作为其特征值。

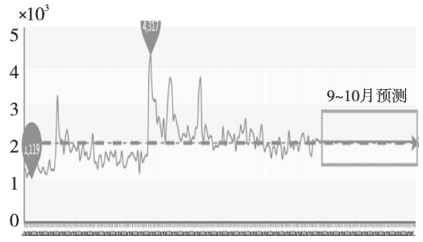
针对 8 月中下旬某些艺人播放量突然成倍增加的情况,采用 E-TSMP 算法中的子序列模式匹配方法进行预测,如图 8 所示,在回落的过程中,不同分段中预测值的变化以梯度回落方式来体现,为了对均值预测的精确度进行改进,使预测值更符合播放量的实际变化趋势;根据处理后的 3 个最佳匹配子序列的样本序列的变化率求得。

第 2 赛季排行榜 | 第 1 赛季排行榜

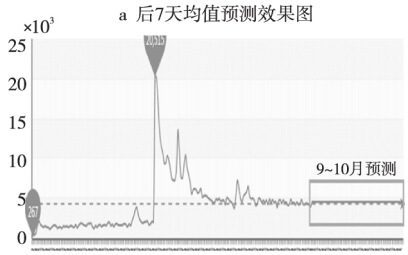
排名	参赛者	所在组织	评分	最优成绩提交日
1	datahacker	南京邮电大学	505151	2016-07-15
2 ↑1	陕西师范大学网络信...	陕西师范大学	501569	2016-07-15
3 ↓1	Heal the World	东北大学	501532	2016-07-15
4 ↑1	data_coders	中国科学院	500916	2016-07-15
5 ↑44	承泽	上海嘉竹唐思金融信息服务有限	500838	2016-07-15
6 ↓2	COM	中国科学院	500526	2016-07-12
7 ↑1	What's matter?	东北大学	499885	2016-07-15
8 ↓1	cike	华南理工大学	499396	2016-07-15
9 ↓3	以手推松	西安交通大学	499390	2016-07-15
10 ↑2	感觉萌萌哒	其它-	499025	2016-07-15

Figure 6 Competition ranking

图 6 比赛排名



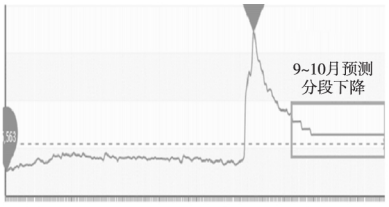
a 后7天均值预测效果图



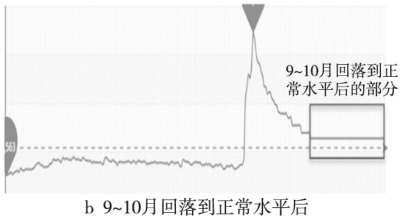
b 后3天均值预测效果图

Figure 7 Experiment results of the category optimal value selection method

图 7 采用类别最优值选择法的实验结果



a 9~10月预测分段下降



b 9~10月回落到正常水平后

Figure 8 Experiment results of the sub-sequence pattern matching method

图 8 采用子序列模式匹配法的实验结果

通过对训练集中的数据进行分析发现,当艺人有新专辑发布时,在随后的 20～30 日内其播放量会有明显的提升。针对 9～10 月有新专辑发布的艺人,在 E-TSMP 算法中对此类别的预测值做了附加处理。如图 9 所示,艺人在 9 月有新专辑发布时,则对发布专辑后一段时间的日播放量有相应提升。

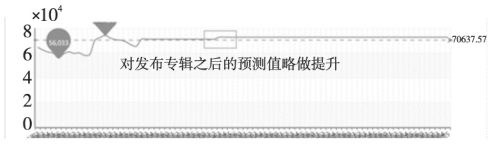


Figure 9 Experiment result of additional processing for releasing new album

图 9 针对新发布专辑进行附加处理的实验结果

在比赛期间曾对所有艺人做过周期性的判断和叠加处理,采用周作为提取和叠加的周期,预测结果如图 10 所示,叠加后导致评判分数有所下降,即降低了预测值与实际播放量的准确度,分析原因如下:

(1)周期性是根据每个艺人样本集的时间序列的最后 10 周 70 天的日播放量进行提取的,提取出来的周期性在未来 60 天的预测结果中不一定存在。

(2)周期性是在预测趋势的基础上进行叠加的,若预测趋势的偏差过大,进行周期性叠加的结果可能会适得其反。

故最终提出的 E-TSMP 算法未对所有艺人的预测值进行周期性的叠加。

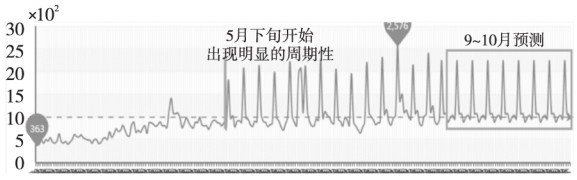


Figure 10 Experiment result of periodic overlay

图 10 周期性叠加的实验结果

5 结束语

在大数据环境下针对音乐的流行趋势预测,本文借鉴了 STL、Holt-Winters 模型中的分解思想,在基于类别最优值选择和子序列模式匹配等方法以及对艺人近期发布新专辑的附加处理的基础之上,提出 E-TSMP 算法。基于阿里云的 Map-Reduce 框架,在阿里音乐平台上实现了对 2016 年 9 月~10 月艺人播放量较为准确的预测,并使用 Echarts 工具实现报表的可视化,直观形象地展示了艺人日播放量序列变化及预测值。

参考文献:

[1] Brockwell P J, Davis R A. Theory and application of time series analysis[M]. 2nd Edition. Tian Zheng, translation. Beijing: Higher Education Press, 2001. (in Chinese)

[2] Wang Yan. Application of time series analysis [M]. 2nd Edition. Beijing: Renmin University of China Press, 2008. (in Chinese)

[3] Wang Zhen-long. Time series analysis [M]. Beijing: China Statistics Press, 2010. (in Chinese)

[4] Wu Huai-yu. Time series analysis and synthesis [M]. 1st Edition. Wuhan: Wuhan University Press, 2004. (in Chinese)

[5] Wu Xi. Time series modeling and application of model selection [D]. Hefei: Hefei University of Technology, 2006. (in Chinese)

[6] He Shu-yuan. Application of time series analysis [M]. 1st Edition. Beijing Peking University Press, 2003. (in Chinese)

[7] Yang Shou-zi, Wu Ya, Xuan Jian-ping, et al. Time series analysis of engineering applications [M]. 2nd Edition. Wuhan: Huazhong University of Science and Technology Press, 1991. (in Chinese)

[8] Li Gui-bin. Comparison of various estimates of differencing order in an ARIMA model [J]. Chinese Journal of Applied Probability and Statistics, 1994, 10(4): 353-362. (in Chinese)

[9] Chatfield C, Yar M. Holt-Winters forecasting: Some practical issues[J]. The Statistician, 1988, 37(2): 129-140.

[10] Cleveland R B, Cleveland W S, McRae J E, et al. STL: A seasonal-trend decomposition procedure based on loess[J]. Journal of Official Statistics, 1990, 6(1): 3-73.

[11] Jain L C, Medsker L R. Recurrent neural networks: Design and applications[M]. 1st Edition. Boca Raton: CRC Press, 1999.

[12] Likas A, Vlassis N A, Verbeek J J. The global *k*-means clustering algorithm[J]. Pattern Recognition, 2003, 36(2): 451-461.

[13] Lai J Z C, Tsung-Jen H. Fast global *K*-means clustering using cluster membership and inequality[J]. Pattern Recognition, 2010, 43(5): 1954-1963.

[14] Yu Hong-yan. Excel statistical analysis and decision-making [M]. Beijing: Higher Education Press, 2001. (in Chinese)

附中文参考文献:

[1] Brockwell P J, Davis R A. 时间序列的理论与方法[M]. 第 2 版. 田铮,译. 北京:高等教育出版社,2001.

[2] 王燕. 应用时间序列分析[M]. 第 2 版. 北京:中国人民大学出版社,2008.

[3] 王振龙. 时间序列分析[M]. 第 2 版. 北京:中国统计出版社,2010.

[4] 吴怀宇. 时间序列分析与综合[M]. 第 1 版. 武汉:武汉大学出版社,2004.

[5] 吴喜. 时间序列建模与模型选择的应用研究[D]. 合肥:合肥工业大学,2006.

[6] 何书元. 应用时间序列分析[M]. 第 1 版. 北京:北京大学出版社,2003.

[7] 杨叔子,吴雅,轩建平,等. 时间序列分析的工程应用[M]. 第 2 版. 武汉:华中理工大学出版社,1991.

[8] 李贵斌. ARIMA 模型差分阶的估计方法的比较[J]. 应用概率统计,1994,10(4):353-362.

[14] 于洪彦. Excel 统计分析与决策[M]. 北京:高等教育出版社,2001.

作者简介:

郁伟生(1972-),男,江苏涟水人,硕士,高级工程师,研究方向为网络大数据分析。**E-mail**:yyy@snnu.edu.cn

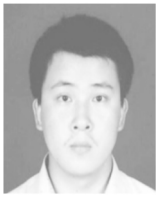
YU Wei-sheng, born in 1972, MS, senior engineer, his research interest includes network big data analysis.

邓伟(1983-),男,陕西渭南人,硕士,工程师,研究方向为网络大数据分析。**E-mail**:vvdeng@snnu.edu.cn

DENG Wei, born in 1983, MS, engineer, his research interest includes network big data analysis.

张瑶(1992-),女,山东庆云人,硕士,研究方向为网络大数据分析。**E-mail**:517789441@qq.com

ZHANG Yao, born in 1992, MS, her research interest includes network big data analysis.



李蜀瑜(1978-),男,四川资中人,博士,副教授,研究方向为网络大数据分析。**E-mail**:lishuyu@snnu.edu.cn

LI Shu-yu, born in 1978, PhD, associate professor, his research interest includes network big data analysis.