基于拐点的网络舆情预测研究

郑步青[1] 邹红霞[2] 胡欣杰[2]
(航天工程大学研究生管理大队 北京 101416)[1] (航天工程大学信息装备系 北京 101416)[2]

摘 要 舆情预测是实现网络舆情监控最重要的一个环节,针对舆情演化过程中的拐点会影响舆情预测的情况,在 ARIMA 和灰色预测的基础上,提出了一种基于拐点的预测方法,建立了分段和镜像处理的数学模型。最后用实例对模型进行对比验证,并总结了模型的优缺点。实验表明,该方法能够减小拐点的影响,提高舆情预测的准确度。

关键词 网络舆情,ARIMA,灰色预测,拐点
中图法分类号 G202 **文献标识码** A

Research on Public Opinion Prediction Based on Inflection Point

ZHENG Bu-qing[1] ZOU Hong-xia[2] HU Xin-jie[2]
(Company of Postgraduate Management, Space Engineering University, Beijing 101416, China)[1]
(Department of Information Equipment, Space Engineering University, Beijing 101416, China)[2]

Abstract Public opinion prediction is an important part of monitoring. In view of public opinion inflection point in the evolution process will affect public opinion forecast, based on ARIMA and gray prediction model, a prediction method based on inflection point was proposed, the mathematical model of segmentation and mirror processing was established. Finally, an example was used to verify the model, and the advantages and disadvantages of the model were summarized. Experiments show that this method can reduce the influence of inflection point and improve the accuracy of public opinion prediction.

Keywords Public opinion, ARIMA, Gray forecast, Inflection point

随着网络的迅速发展,对舆情的监控预测显得越来越重要。学者们对舆情的预测做了较多研究,为提高预测准确度,提出了不同的模型。文献[1]以分解舆情演化过程为思路,对其进行 EMD 分解,得到演化的趋势、周期、突发和随机成分,以提高预测效果;文献[2]基于网络舆情发展过程的复杂性和多成分的特点,提出一种基于小波分析和人工神经网络的网络舆情建模和预测方法;文献[3]针对传统预测方法无法有效预测 Web 舆情的长期趋势中的拐点,提出了一种 Web 舆情长期趋势预测方法,弥补了传统方法无法预测拐点的缺陷[3]。这些预测研究大部分都是针对舆情演化中的特点进行舆情的预测,同样地本文针对舆情演化过程的拐点的特性,研究了拐点对舆情预测的影响,判断了舆情"拐点"的存在位置,提出了基于 ARIMA 和灰色预测的分段和镜像处理模型,提高了预测的准确率。

1 ARIMA 与灰色预测模型

ARIMA 和灰色预测模型都是预测中常用的模型,在 ARIMA 模型中,时间序列变量被认为是变量的过去观测值和随机干扰误差的线性函数,通过建立的数学模型来预测未来值;而在灰色预测中,则通过对原始数据进行处理后建立相应的微分方程模型来进行预测,对小样本的预测效果较好。这两种预测模型的共同点在于本质上只能捕捉线性关系,而不能捕捉非线性关系。

1.1 ARIMA 预测模型

ARIMA 模型的全称叫做自回归移动平均模型,记作 $ARIMA(p,d,q)$,是统计模型中最常见的一种用来进行时间序列预测的模型。ARIMA 模型的基本思想是:将预测对象随时间推移而形成的数据序列视为一个随机序列,用一定的数学模型来近似描述这个序列。这个模型一旦被识别后就可以根据时间序列的过去值及现在值来预测未来值。$ARIMA(p,d,q)$模型的实质是先对非平稳的历史数据 y_t 进行 d 次差分处理,得到新的平稳的数据序列 X_t,将 X_t 拟合 $ARMA(p,q)$模型,然后再将原 d 次差分还原,便可以得到 y_t 的预测数据。

自回归移动平均模型 $ARMA(p,q)$的一般表达式为[4]:

$$X_t=\varphi_1 X_t+\cdots+\varphi_p X_p+\varepsilon_t+\beta_1\varepsilon_{t-1}+\beta_2\varepsilon_{t-2}+\cdots+\beta_q\varepsilon_{t-q} \tag{1}$$

当 $q=0$ 时,该模型成为自回归 $AR(p)$模型:

$$X_t=\varphi_1 X_t+\cdots+\varphi_p X_p+\varepsilon_t \tag{2}$$

当 $p=0$ 时,该模型成为移动平均 $MA(q)$模型:

$$\varepsilon_t+\beta_1\varepsilon_{t-1}+\beta_2\varepsilon_{t-2}+\cdots+\beta_q\varepsilon_{t-q} \tag{3}$$

其中,前半部分为自回归部分,非负整数 p 为自回归阶数,φ_1,φ_2,…,φ_p 为自回归系数,后半部分为滑动平均部分,非负整数 q 为滑动平均阶数,β_1,β_2,…,β_q 为滑动平均系数,ε_t 是误差项。

目前 ARIMA 模型由于其简单性、可行性和灵活性,已经

郑步青(1993—),男,硕士生,主要研究方向为舆情数据的处理;**邹红霞** 女,副教授,主要研究方向为信息对抗和信息处理;**胡欣杰** 女,教授,主要研究方向为信息安全。

成为时间序列分析预测最常用的一种方法。模型的优点主要有模型十分简单,只需要内生变量而不需要借助其他外生变量,而缺点是要求时序数据是稳定的,或通过差分后是稳定的[5]。

1.2 灰色预测

灰色预测是一种对含有不确定因素的系统进行预测的方法。灰色预测通过鉴别系统因素之间发展趋势的相异程度,即进行关联分析,并对原始数据进行生成处理来寻找系统变动的规律,生成有较强规律性的数据序列,然后建立相应的微分方程模型,从而预测事物未来的发展趋势。一切灰色序列都能通过某种生成弱化其随机性,显现其规律性。为了弱化原始时间序列的随机性,一般在建立灰色预测模型之前会对序列进行数据处理,数据处理的常用方式有累加生成和累减生成。

1)累加生成:将原始序列通过累加得到生成列。它的规则是将第一个数据作为生成列的第一个数据,将原始序列的第二个数据加到原始序列的第一个数据上,其作为生成列的第二个数据。以此类推,即可得到生成列。

记原始时间序列为:

$$X^{(0)}=\{X^{(0)}(1),X^{(0)}(2),\cdots,X^{(0)}(n)\} \tag{4}$$

生成列为:

$$X^{(1)}=\{X^{(1)}(1),X^{(1)}(2),\cdots,X^{(1)}(n)\} \tag{5}$$

2)累减生成:将原始序列前后两个数据相减得到累减生成列。累减是累加的逆运算,可将累加生成列还原成非生成列,在建模中获得增量信息。

一次累减的公式为:

$$X^{(1)}(k)=X^{(0)}(k)-X^{(0)}(k-1) \tag{6}$$

灰色预测具有不需要大量数据,能解决历史数据匮缺、序列不够完整和数据可靠性低等问题的特点,是处理小样本预测问题的有效途径,且具有运算简单、易于检验和不必考虑分布规律的特点,但其对波动性大的序列预测结果较差[6]。

2 基于拐点的预测模型

本文针对舆情拐点所带来的预测不精确的问题,首先对拐点的节点位置进行了判断,然后以灰色预测和 ARIMA 模型为基础,讨论了针对拐点的两种预测模式,分别为分段处理和镜像处理,最后进行误差比较。预测流程如图 1 所示。

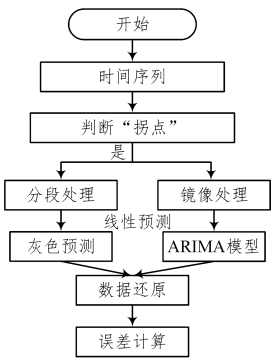

图 1 基于拐点的预测流程图

2.1 拐点定义及判断

1)拐点定义

目前已经有较多文章对舆情拐点进行了研究,文献[7]认为拐点是趋势拐点,分割了舆情演化阶段;文献[8]认为过去位次较高或较低的舆情信息在一段时间内关注度急速下降或者上升,与之前的一段时间变化趋势相比形成转折的点叫做拐点。拐点在舆情中是指趋势转折点,在拐点前后舆情的演化趋势不一致,序列变得非线性化、不平稳,导致舆情预测产生较大的误差,主要有阶段拐点和极值拐点。阶段拐点是区别舆情演化 3 个阶段——潜伏阶段、爆发阶段、消亡阶段的重要节点。同样,舆情演化的极值点,即极大值点,存在着上升趋势变为下降趋势的变化,也是研究舆情演化的重要节点。结合趋势性分析和对拐点的判断可以得到舆情演化阶段的信息,为后期的预测和建模提供分段处理的理论基础。

2)拐点的判断

高辉等[3]指出传统预测方法无法有效预测 Web 舆情的长期趋势中的拐点,提出了一种 Web 舆情长期趋势预测方法,弥补了传统方法无法预测拐点的缺陷。陈福集等[7]提出了一种基于 E-divisive 的时间序列阶段定量划分趋势拐点的方法,通过判断时间序列段的分布是否相同来定位趋势转折点,再通过增加时间节点进行迭代定位,估计出所有的趋势转折点,获得不同趋势的时间序列片段。

本文采用 E-Divisive 方法对拐点进行了判断,E-Divisive 方法是一种检测时间序列趋势拐点的算法,它不需要设定前提条件,通过动态检测的算法,将复杂的时间序列数据分段以确定趋势变化,但不适用大样本的计算[9]。它通过迭代地应用定位单个变化点的过程来估计多个变化点,在每次迭代中估计新的变化点位置,以便分割现有的分段。

变化点的估计是通过置换检验确定的,输入时间序列 $y_1,y_2,\cdots,y_n$,检验水平的最小值 $P_0=0.05$。假设在第 k 次迭代中当前的变化点将时间序列分为 k 个部分:$S_1,S_2,\cdots,S_k$,并且估计下个变化点为 $\hat{T}_k$,相关的检验统计值为 q_0。通过交换观测值获得交换样本,在交换样本上估计下个变化点的位置 $\hat{T}_{k,r}$,与其相关的检验统计值为 q_r,R 是排列总数,我们近似地估计 $\hat{p}$ 值为:

$$\hat{p}=\#\frac{\{r:q_r\geqslant q_0\}}{R+1} \tag{7}$$

如果估计的 $\hat{p}$ 值小于 P_0,则存储为新的变化点,在新的时间段内继续寻找下一个变化点,继续迭代计算;如果估计的 $\hat{p}$ 值大于 P_0,则拒绝该假设,该时间序列段无新的变化点,对下一个时间段进行置换检测。这种方法的时间复杂度为 $O(kT^2)$,其中 k 表示估计的变化点的个数,T 表示被观测的时间序列的总数。我们通过引 R 包 ecp 进行多变量时间序列的多变点分析以实现算法。ecp 软件包提供了变化点分析的方法,能够检测出任何类型的时间序列的分布变化和变化点的数量和位置。

2.2 拐点预测方法

1)分段预测

目前大部分文献对舆情拐点的处理都是通过分段来降低拐点带来的影响。文献[10]提出了一种序列可变长度的机制,其实质在于在预测过程中通过预测结果的精确度确定拐点是否存在,若预测拟合不好,则预测序列变为原序列的一半。本文首先确定了拐点的位置,然后对拐点前后不同的趋势序列分别进行灰色预测,将得到的两段预测结果组合成预测序列,再与原始序列进行误差计算。一般来说,一段舆情具有两个阶段拐点和一个极值拐点,故可将舆情分为 4 个部分:

1)舆情的潜伏阶段；2)爆发阶段拐点到极值拐点；3)极值拐点到消亡阶段拐点；4)消亡阶段。

2)镜像处理

本文提出的镜像处理是针对单调递增或递减序列经过极值拐点之后趋势发生变化的情况，目的是为了减少拐点带来的非线性影响，增加了时间序列的线性程度，结合 ARIMA 预测模型的性质，使得预测更加准确。采用镜像处理的条件是极值拐点两边的变化率是对称的，否则经过镜像处理后的数据平稳性不高。

镜像预测步骤如下：

①根据极值拐点的判断，确定舆情时间序列(y_1，y_2，…，y_n)变化的趋势(单调递增变为单调递减)。

②镜像变换，以极值拐点处为对称轴 $y=y_a$，对极值拐点后的数据进行镜像处理得到序列(y_1'，y_2'，…，y_n')=(y_1，…，y_a，y'_{a+1}，…，y_n')。

③线性预测，对得到的新序列 y_1'，y_2'，…，y_n'进行 ARIMA 建模预测。

④数据还原，对预测后得到的数据进行对称轴还原，得到镜像预测数据$\bar{y}_1$，$\bar{y}_2$，…，$\bar{y}_n$。

⑤误差分析，预测数据与原始数据进行比较处理，得到误差。

2.3　模型误差分析

基于“拐点”的预测方法能够找到舆情的拐点，减少了拐点带来的影响，同时在一定程度上减少了非线性影响因素带来的影响，能够提高预测精确度。为了更直观地对比各改进模型的预测效果，引入 MSE(均方误差)和 MAE(绝对评价误差)两种指标来评估模型，其中，N 为预测样本序列总数量，y_i 是原始数据，$\tilde{y}_i$ 是预测值。

$$MSE=\frac{1}{N}\sum_{1}^{N}(y_i-\tilde{y}_i)^2 \tag{8}$$

$$MAE=\frac{1}{N}\sum_{1}^{N}|y_i-\tilde{y}_i| \tag{9}$$

3　实验验证

本文将文献[11]中的对 2013 年引起全民关注的“刘志军案”进行处理后得到的热度值作为特征量进行时间序列分析，基于文献中的数据进行分析更具有代表性和准确性。表 1 列出了舆情在不同时间片内的热度值。

表 1　舆情在不同时间片内的热度值

时间/天	1	2	3	4	5	6	7	8
热度值	40	90.6	196	388.4	663.4	935.6	1100.6	1135.8
时间/天	9	10	11	12	13	14	15	16
热度值	1084.4	990.6	884.2	778.8	681	592.8	514.4	445.6

1)拐点的判断

提取原始序列 y_i，原数据个数为 16 个，进行拐点分析时数据明显不够。首先用 MATLAB 进行插值处理，在每个时间节点内插入 9 个数据，总得到 151 个数据。再通过 R 语言编程调用 E-Divisive 函数对时间序列进行判断，可以得到图 2 的阶段划分图。图 2 中虚线就是拐点的位置，即 $t'=40$，$t'=110$，$t'=152$ 的节点。由于 $t'=152$ 是不存在的节点，因此不考虑。$t'=40$ 对应的原时间节点为 $t=4$，$t'=110$ 对应的原时间节点为 $t=11$，即可以得到阶段拐点的节点位置。

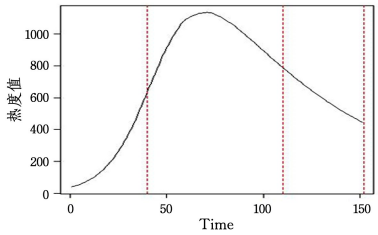

图 2　舆情拐点图

2)分段处理预测

我们在步骤 1)中求出了阶段拐点的位置，且极值拐点在 $t=8$ 时达到最大值。故该舆情可分为 4 段：第一段，$t=1-4$，潜伏阶段；第二段，$t=5-8$，爆发前半阶段；第三段，$t=9-11$，爆发后半阶段；第四阶段，$t=12-16$，消亡阶段。

图 3 是对舆情原始时间进行分段灰色预测后得到的组合预测曲线图，整体上对比分析预测的准确度较高，在上升阶段的预测误差较大，在下降阶段的灰色预测较小。

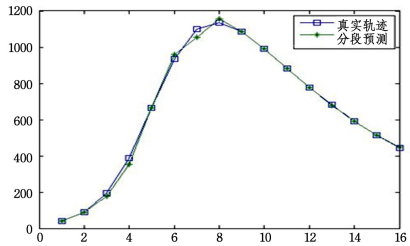

图 3　灰色分段预测对比图

3)镜像处理预测

首先对热度的变化率进行计算可以发现，其在极值拐点前后的变化率相差不大，可以对其进行镜像处理，处理后得到热度镜像曲线图，如图 4 所示。

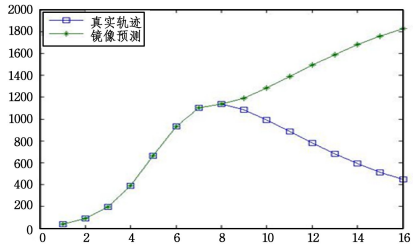

图 4　镜像数据图

将原始数据和镜像数据进行 ARIMA 预测，原始数据的 ARIMA 模型为(2,4,3)，镜像数据得到的 ARIMA 模型为(1,2,1)，通过对比可以发现镜像处理后得到的数据较原始数据，线性程度有所提高，平稳性有所提高，预测所得到的数据也较为准确。镜像预测对比图如图 5 所示。

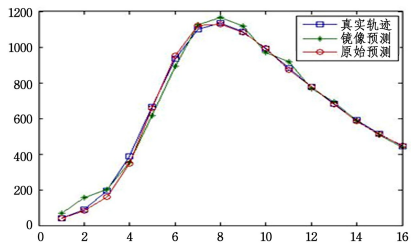

图 5　镜像预测对比图

4)预测误差结果对比分析

对各改进组合模型进行 MSE(均方误差)、MAE(绝对评

(下转第 575 页)

表 CPU 中的点偏移 MC 改进算法。从表 1 可以看出,所有在 CUDA 架构下的并行计算都极大地提高了执行速度,基本可以实现实时改变阈值的交互式建模。

表 1　实验数据表

(单位:ms)

算法 \ 阈值	$\beta_0=45$	$\beta_1=31$	$\beta_2=25$	$\beta_4=8$
a	13	13	15	16
b	219	560	900	980
c	3195	4245	5054	6776

结束语　本文结合 CUDA 架构的可编程性和 Marching Cubes 算法分而治之的特点,合理设计核函数,并行处理 MC 算法中活跃体素以及活跃边的提取,加快执行速度。同时,引入中点投影的方式,改善活跃边与等值面平行的状态,从而达到提高网格质量的目的,研究表明,本文算法的实验效果较好。

但该算法有两点不足之处,首先是投影方法的迭代次数由人为确定,并且只能在 CPU 中进行,算法的大部分时间都集中在这个部分;其次由于硬件条件,该算法无法直接读入数据量稍大的体数据。针对以上两点,需要后续进行进一步的研究。

参 考 文 献

[1] LORENSEN W,CLINE H. Marching Cubes:A High Resolution 3DSurface Construction Algorithm [J]. Computer Graphics (S0097-8930),1987,21(4):163-169.

[2] 朱恺. 基于改进 MC 算法的脑图谱三维可视化应用研究 [D]. 太原:太原理工大学,2015.

[3] CHANG M,WOONG O J,CHANG D S,et al. Interactive marching cubes algorithm for intraoral scanners [J]. The International Journal of Advanced Manufacturing Technology,2017,89(5):2053-2062.

[4] 周筠,樊晓平,蒋富. 医学仿真中一种高效的生物组织几何建模方法 [J]. 系统仿真学报,2012,24(1):6-10.

[5] SUN L N,TIAN H Q,WU D M,et al. Three-Dimensional Geometric Modeling of the SpineBased on Reverse Engineering Technology [C]// 3rd International Conference on Biomedical Engineering and Informatics. 2010:1292-1295.

[6] 王明,冯洁青,杨贲. 移动立方体算法与移动四面体算法的对比与评估 [J]. 计算机辅助设计与图形学学报,2014,26(12):2009-2106.

[7] CIZNICKI M,KIERZYNKA M,KUROWSKI K,et al. Efficient Isosurface Extraction Using MarchingTetrahedra and Histogram Pyramidson Multiple GPUs [C]// International Conference on Parallel Processing and Applied Mathematics. 2011:343-352.

[8] RECK F,DACHSBACHER C,GROSSO R,et al. Realtime isosurface extraction with graphics hard-ware [R]. Eurographics Short Presentations,2004.

[9] 汤颖,嵇海锋,盛风帆,等. 大规模森林多精度生长仿真模型及其计算加速算法[J]. 小型微型计算机系统,2016,37(5):1033-1038.

[10] HAN S Q,LEI Z,SHEN W F,et al. An Approach to Improving the Performance of CUDA in Virtual Environment [C]// IEEE/ACIS International Conference on Software Engineering,Artificial Intelligence,Networking and Parallel/Distributed Computing (SNPD). 2016:585-590.

[11] DIETRICH C A,SCHEIDEGGER C E,SCHREINER J,et al. Edge transformations for improving mesh quality of marching cubes [J]. IEEE Transactions on Visualization and Computer Graphics,2009,15(1):150-159.

[12] DIETRICH C A,SCHEIDEGGER C E,COMBA J L D. Edge groups:an approach to understanding the mesh quality of marching methods[J]. IEEE Transactions on Visualization and Computer Graphics,2008,14(6):1651-1666.

(上接第 541 页)

价误差)两种指标的计算,结果取两位有效数字,如表 2 所列。

表 2　各模型的预测误差对比结果

	原始序列	分段处理	镜像处理
MSE	14.35	4.38	7.68
MAE	45.63	9.52	25.86

从结果中可以看出,分段处理和镜像处理都比原始序列直接预测的结果更为精确,说明两种处理方法都能够有效地减小误差,提高预测的精确度。其中分段预测的结果最为准确,镜像处理后使得非线性误差变大,因此预测效果不如分段处理。

结束语　本文针对舆情预测中拐点存在的造成预测不准确的问题,首先判断了舆情拐点存在的位置,提出了基于 ARIMA 和灰色预测的分段和镜像处理模型,对分段处理和镜像处理分别进行了研究和讨论,最后用实例验证了分段和镜像处理能够在一定程度上提高预测的准确率。今后在舆情预测上可以考虑优化非线性部分,减小非线性误差。

参 考 文 献

[1] 周耀明,王波,张慧成. 基于 EMD 的网络舆情演化分析与建模方法[J]. 计算机工程,2012,38(21):5-9.

[2] 舒予,张黎俐. 基于小波分析与人工神经网络的网络舆情预测[J]. 情报科学,2016,34(4):40-42.

[3] 高辉,王沙沙,傅彦. Web 舆情的长期趋势预测方法[J]. 电子科技大学学报,2011,40(3):440-445.

[4] 何炎祥,刘健博,孙松涛. 基于神经网络的微博舆情预测方法[J]. 华南理工大学学报(自然科学版),2016,44(9):47-52.

[5] 王努努,张伟佳,钮亮. 基于 ARIMA 和 BP 神经网络模型的舆情情感预测[J]. 电子科技,2016,29(5):83-87.

[6] 杜智涛,谢新洲. 利用灰色预测与模式识别方法构建网络舆情预测与预警模型[J]. 图书情报工作,2013,57(15):27-33.

[7] 陈福集,张燕. 基于 E-Divisive 的网络舆情演化分析[J]. 情报杂志,2016,35(4):75-79.

[8] 黄惠新,陈越,李超零,等. 基于 OLAP 技术的网络舆情分析[C]// 河南省计算机学会 2011 年学术年会. 2011.

[9] JAMES N A,MATTESON D S. ecp:An R Package for Nonparametric Multiple Change Point Analysis of Multivariate Data [J]. Journal of Statistical Software,2013,62(7):1-25.

[10] 荣自瞻. 网络舆情预警与预案系统的研究与实现[D]. 北京:北京邮电大学,2013.

[11] 陈婷,曲霏,陈福集. 基于时间片划分的舆情话题演化模型研究[J]. 武汉:华中师范大学学报(自然科学版),2015,49(6):890-894.