基于大数据的时间序列预测研究与应用

程艳云,张守超,杨 杨
(南京邮电大学 自动化学院,江苏 南京 210023)

摘 要:针对传统时间序列预测算法在分析海量数据时预测精度与预测速率低下的问题,提出一种全新的时间序列预测算法,研究如何将大数据技术应用到移动通信网时间序列形式的核心性能指标(KPI)预测中。文中首先介绍了移动通信网性能指标预测的意义及传统时间序列预测算法的缺陷。其次,基于移动通信网及时间序列特性,给出了基于大数据的时间序列预测算法的理论推导过程,通过大数据方法将时间序列分解为四个不同分量并进行特征提取,根据提取结果进行预测分析。最后,介绍了方法的实现过程,采用真实网络核心性能指标进行实验对比分析,验证该方法的可行性与效率。实验结果表明,基于大数据的时间序列预测算法相比于传统的时间序列预测算法,具有更高的预测精度、更快的预测速率。

关键词:大数据;时间序列;预测分析;移动通信
中图分类号:TN915.07 **文献标识码**:A **文章编号**:1673-629X(2016)06-0175-04
doi:10.3969/j.issn.1673-629X.2016.04.039

Research and Application of Time Series Forecasting Based on Big Data

CHENG Yan-yun, ZHANG Shou-chao, YANG Yang
(College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: According to the detection accuracy and efficiency limitation of traditional time series forecasting methods when dealing with a large amount of data, a new time series forecasting method is put forward to study how to apply the big data technology into Key Performance Index (KPI) prediction of mobile communication network, which is form of time series. First, it introduces the significance of KPI prediction for mobile communication network and the defects of traditional time series prediction algorithm in this paper. Secondly, the theoretical derivation of time series prediction algorithm based on the big data is presented according to the characteristics of mobile communication network and time series. The time series is decomposed into four different components and the feature is extracted by the big data method, and the forecasting analysis is carried out according to the results of the extraction. Finally it gives implementation process and uses the real network KPI to carry out experimental comparative analysis for verification of the feasibility and efficiency of the big data method. The experimental results show that the big data method has higher precision and rate compared with traditional methods.

Key words: big data; time series; forecasting analysis; mobile communication

0 引 言

通信网络中的各项核心性能指标[1](KPI)的预测分析对于通信网络优化至关重要,而通信网络中的各项 KPI 一般均以时间序列形式[2]表示。传统的时间序列分析预测方法包括 Holt-Winters[3]、ARIMA[4]、AR、MA、Vector Auto Regression、梯度回归等。然而,传统的通信网性能预测分析所选用的数据量很小且缺乏实时性,实验结果的准确率也有待提高,而且随着时间的推移,通信网络中的数据量越来越大。到 2020 年,全球以电子形式存储的数据量将达 35 ZB,是 2009 年全球存储量的 40 倍[5]。如此大的数据量,传统的数据库工具无法负担,必须采用专用数据挖掘与分析工具进行分析处理。不过,尽管这些数据挖掘工具价格昂贵,挖掘效果却仍有待提高。

因此,必须采用新的方法来解决这一问题。文中提出的基于统计模型的大数据算法分析利用真实的测量数据而不是模拟仿真数据或假设场景来研究无线网络的预测问题。文中首先利用统计模型对海量数据进

收稿日期:2015-06-28 **修回日期**:2015-10-13 **网络出版时间**:2016-03-22
基金项目:江苏省自然科学基金(BK20140877,BE2014803)
作者简介:程艳云(1979-),女,副教授,硕士生导师,从事自动控制原理、网络优化的教学科研工作;张守超(1991-),男,硕士研究生,研究方向为大数据挖掘在通信网络中的应用。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160322.1518.040.html

行分类处理,并进行特征提取,区分小区类别,然后采用大数据算法分析海量实时数据,并对建立的模型进行参数优化,最终得到预测模型。

1 大数据算法分析

时间序列预测算法主要包括趋势分量预测、季节性分量预测、突发分量预测以及随机误差分量预测。以传统的时间序列预测算法为例,Holt-Winters 算法中 α,β,γ 分别为水平项、趋势项、周期项的平滑参数。由于 α,β,γ 一旦确定就不可以改变,且需要反复试验确定最佳值,因此传统的 Holt-Winters 算法对于长期大量的数据分析是不适合的[6]。而 ARIMA 仅在短期预测中有较好的预测结果,随着预测时间的推迟,其预测误差会越来越大[7],因此 ARIMA 对于长期数据预测是不符合要求的。文献[8-9]对 Holt-Winters 进行了一些改进,文献[10]对 ARIMA 进行了一些改进,但是对于海量数据的长时间预测效果,其结果仍然不符合要求,所以必须采用新的时间序列预测模型来进行预测分析。

文中提出的大数据算法采用全新的方法来对四个分量进行预测。利用海量数据的优点,将隐藏在数据背后的有效信息挖掘出来,具体推导过程如下所示:

(1)趋势分量 $T(t)$ 的预测。

将每一段的起始无线网络话务量历史数据 X_k 和斜率 Slope_k 拟合为一条直线,每个拟合线间首尾连续,将无线网络话务量历史数据作为训练样本进行建模,获得趋势分量 $T(t)$ 预测模型:

$$T(t)=f\{X_k,\text{Slope}_k\}$$

$$K'_{T+i}=\text{Max}\{K_{T+i},\gamma\cdot\min\{K_T,K_{T-1},\cdots,K_{T-N+1}\}\} \tag{1}$$

式中,K'_{T+i} 表示补偿后的改善斜率,如果最近连续 N 个斜率不小于零,那么第 $N+1$ 的斜率不应小于零;γ 是可调节的,直到一个最佳常数。

图 1 展示了趋势分量预测过程。

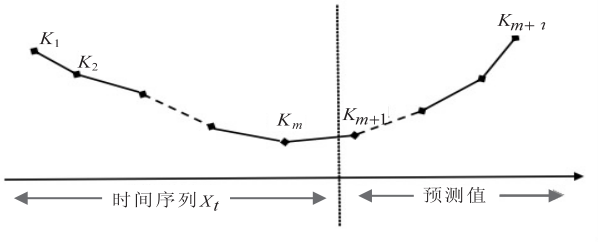

图 1 趋势分量预测过程

如此一来,对于趋势分量 $T(t)$ 的预测,预测值之前数据的权重变成由历史数据 X_k 和斜率 Slope_k 决定。

(2)季节性分量 $S(t)$ 的预测。

文中首先需要确认的是周期时间,通过统计分析对海量数据进行特征提取,按照式(2)进行差分运算,得到矩阵 $\boldsymbol{A}$。

$$\boldsymbol{A}=\begin{bmatrix} x_2-x_1 & x_3-x_2 & \cdots & x_{n-1}-x_{n-2} \\ x_3-x_1 & x_4-x_2 & \cdots & x_n-x_{n-2} \\ \vdots & \vdots & \vdots & \vdots \\ x_m-x_1 & x_{m+1}-x_2 & \cdots & x_{n+m-3}-x_{n-2} \end{bmatrix} \tag{2}$$

对矩阵 $\boldsymbol{A}$ 的每一行进行线性拟合,得到不同的拟合直线 $Y=aX+b$,其中拟合误差最小的行数即为周期 L。p 表示每个周期 L 里的样本数,每个 $q(q=1,2,\cdots,p)$ 位置处的季节分量可表示为 p 样本中相同位置 q 处的数据的平均值,利用式(3)可得出季节性分量。

$$S_{pi}(t)=\sum_{q=1}^{p}X_{qi}/p \tag{3}$$

(3)突发分量 B 的预测。

突发分量 B 产生的原因一般是由于突发事件,比如重大节日、活动、会议等。一般情况下,突发分量具有可列举性,即每个小区的 KPI 对应的突发分量 B 都可以用特定的类别对应特定的数值表示,如式(4):

$$\begin{aligned} B(t)_{\text{value}}&=\{\text{Burst}_{v1},\text{Burst}_{v2},\cdots,\text{Burst}_{vn}\} \\ B(t)_{\text{type}}&=\{\text{Burst}_{t1},\text{Burst}_{t2},\cdots,\text{Burst}_{tn}\} \\ B(t)&=\{\text{CELLID},\text{Burst}_{v1},\text{Burst}_{t1},\cdots,\text{Burst}_{vn},\text{Burst}_{tn}\} \end{aligned} \tag{4}$$

在 KPI 分析预测中,只需要根据小区的 ID 号,查找对应的突发分量 $B(t)$ 带入预测公式即可。

(4)随机误差分量 R 的预测。

在大数据预测模型中,随机误差分量不再是独立分布,而是根据无线网络话务量历史数据减去趋势分量、季节性分量和突发分量得到随机误差分量的预估值。处理的结果确保了随机误差分量更具有实际性。

(5)KPI 预测。

预测目标 KPI 时,利用公式 $X(t)=(1+B(t))\times(T(t)+S(t)+R(t))$ 即可得到目标结果。

2 大数据算法在 KPI 预测中的实现

在通信网中,每个 RNC 下包含大量的小区(一般为 500 ~ 1 000),而每个小区的 KPI 又数量众多(一般为 200 个)。以一年时间为周期计算,每个 KPI 每年数据值为 17 520 个,单个 RNC 内所有小区的一年内所有 KPI 总数将过亿。考虑到数据量巨大,采用大数据进行的 KPI 预测分析,需要对小区数据进行一定的处理,具体步骤如图 2 所示。

步骤 1:插值处理。

在数据导入之前,需要对数据进行预处理,处理的主要工作为缺值插入。文中采用的插入方法为构建线

性拟合曲线,具体做法为以缺失值前几点、后几点作为一个数据序列,做一个最小二乘法的线性回归[11],将对应缺失的这点代入线性回归方程,得出缺失点的值。

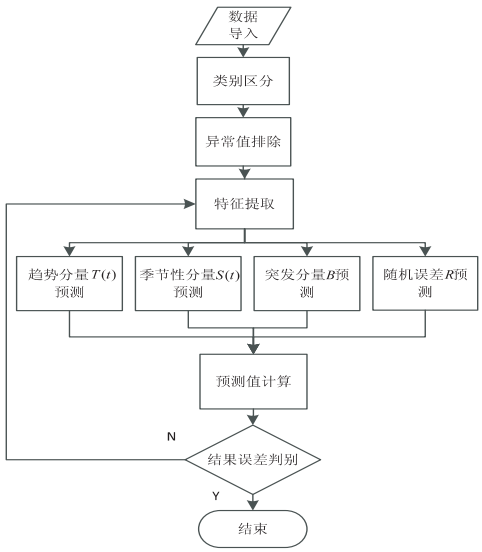

图 2　大数据预测模型流程图

步骤 2:小区分类。

对所有小区进行分类处理,将所有小区的忙时进行特征提取,得到不同忙时的特征,区分出不同类别的小区,然后再对每种类型的小区进行分析预测。小区类别事先未知,文中采用统计方法,将所有 RNC 下所有小区的一天 KPI 特性进行统计分析,得到不同时间分布的忙时,从而得到不同类别的小区。

步骤 3:异常值排除。

对于每种类型数据,取可信度 95%,其边界为 $u-2\sigma$ 和 $u+2\sigma$,来排除异常值。如果时间序列不符合正态分布,则不能通过测试,此时应该采用其他方法来排除异常值。

步骤 4:预测分析。

排除异常值之后,根据特征提取结果确定一维周期值[12-13],利用大数据算法分别进行趋势分量预测、季节性分量预测、突发分量预测及随机误差分量预测。

步骤 5:结果判定。

对于分别预测得到的趋势分量、季节性分量、突发分量以及随机误差分量,通过公式 $X(t)=(1+B(t))\times(T(t)+S(t)+R(t))$ 得到最终预测值,判别与真实值之间误差是否在可接受范围内,若是,则模型建立成功,若否,返回修改模型参数。

3　实验结果

以通信网络中某一性能指标(RRC 设置成功率)为例。首先任取某一小区,采用不用方法分别对该小区的 RRC 设置成功率进行长期预测和短期预测,并对结果进行对比分析;其次,对 RNC 内所有小区进行预测,并对结果进行分析比较。

首先对所有 RNC 内的小区进行分类处理,根据忙时不同特征分布可以区分得到 7 种不同类型的小区。选取某一类型小区的某一小区连续 30 天数据为初始数据集,分别采用不同算法预测不同长度值。先进行周期特征提取,按照式(2)得到矩阵 $\boldsymbol{A}$,并对 $\boldsymbol{A}$ 的每行数据进行线性拟合。对于每条拟合直线,采用最小二乘法计算误差,通过计算得到当 $L=48$ 时,误差最小,即周期为 48。

图 3 展示了 Bigdata 算法对应不同周期 L 的预测结果。其中点代表预测值,线条代表真实值走势,虚线表示初始值与预测值分界线。

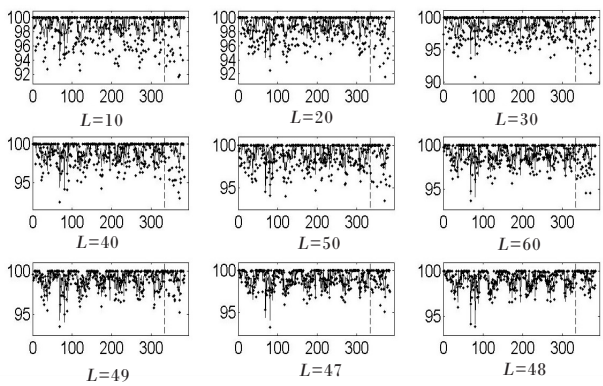

图 3　Bigdata 算法对应不同周期 L 的预测结果

图 4 展示了 RRC 设置成功率的实际值与 Holt-Winters 算法、ARIMA 算法以及基于大数据算法的预测值对比结果。显而易见,基于大数据算法的预测结果与实际值具有很大的重合性。

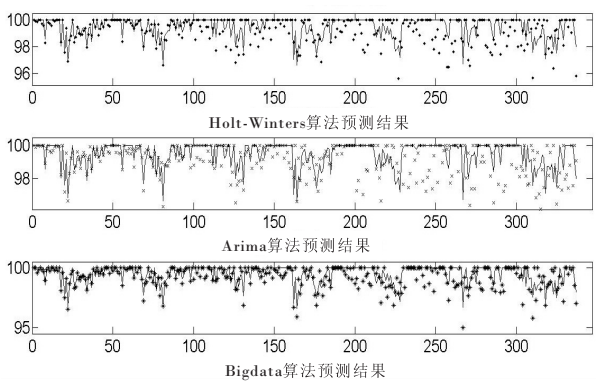

图 4　单小区 RRC 设置成功率预测结果对比图

通过统计计算可以得到,在大数据预测模型中,初始数据预测的平均绝对百分比精度[14](误差结果在 1% 以内)是 95.28%,预测结果平均绝对百分比精度是 90.47%。相比于 Holt-Winters 算法、ARIMA 算法的 78.28% 和 70.1% 均有很大提高。

表 1 展示了 Bigdata 算法、Holt-Winters 算法和 ARIMA 算法三者在长/短期初始数据预测与结果预测精度对比。

通过表中数据可以得到,基于大数据的方法在长

期预测跟短期预测的精度差距很小,尤其在预测结果精度方面,而基于 Holt-Winters 方法和 ARIMA 方法的预测在长期跟短期结果出现大幅度的下降,即基于大数据方法相比于 Holt-Winters 方法和 ARIMA 方法更加适用于长期的时间序列预测。此外,短期预测中三种方法所需时间均在 20 s 内,但是在长期大量数据预测时,基于大数据的方法所需时间仅为另外两种方法的一半,约为 100 s。

表 1 不同方法对应长/短期预测结果对比 %

指标算法	短期初始	短期预测	长期初始	长期预测
Bigdata	95.28	90.47	93.03	90.06
Holt-Winters	90.94	78.28	84.75	75.43
ARIMA	79.33	70.10	70.69	62.88

同样选取商业型小区的某一 RNC 级别内所有小区(共计 478),预测某天(周一)忙时(晚上 8 点)所有小区的性能指标值。图 5 展示了 RNC 内所有小区的实际值与预测值对比,其中点代表预测,线条代表真实值走势。

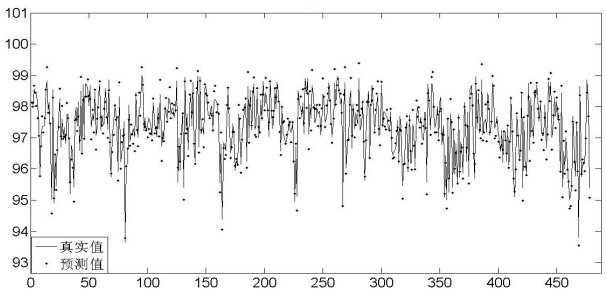

图 5 RNC 级小区 RRC 预测值对比图

在大数据预测模型中,所有小区性能指标的预测值平均绝对百分比精度是 84.66%,高于传统方法的预测精度。

通过分析比较结果可以得出,基于大数据的预测模型的预测结果在长时间预测、大范围预测均能满足要求,相比于传统的预测方法,采用大数据技术的预测模型具有更高的精度以及更快的速度。总体来说,通信网络中的 KPI 都可以通过预测模型得到结果,这两项数值都在可以接受的范围内,并且未来还有提高的空间,尤其对于单小区的长时间预测结果精度。

4 结束语

新颖的大数据技术及其算法可以克服传统网络仿真中的缺点,基于统计模型的大数据算法的无线网络性能分析将使得网络特征、用户特征、话务流量特征等在网络性能分析评估中得到最准确和最真实的反应[15]。文中的大数据算法模型将使得埋藏在海量数据背后的网络行为特征得以准确挖掘出来,从而使得传统的网络性能分析这一领域到达一个新的台阶。

文中仅对网络 KPI 进行预测分析,对于网络优化中的其他问题,还有待进一步的研究,包括:

(1)预测网络话务和流量的短期—长期趋势;

(2)基于网络话务来推测网络容量的变化趋势。

中国从 2013 开始大规模商用 TDD LTE 网络,此方法采用基于大数据的算法分析的网络性能以及质量评估系统,采用实时数据进行预测分析,预测结果也能够满足需求,在未来具有很高的应用前景。

参考文献:

[1] RAN 14.0 KPI 参考手册—2 版[M]. 出版地不详:华为技术有限公司,2012.

[2] 林国华. 时间序列分析法在移动通信数据分析中的研究与应用[D]. 广州:广州工业大学,2013.

[3] Szmit M,Szmit A. Use of holt-winters method in the analysis of network traffic:case study[J]. Communications in Computer & Information Science,2011,160:224-231.

[4] Box G E P,Jenkins G M,Reinsel G C. 时间序列分析:预测与控制[M]. 王成璋,尤梅芳,郝 杨,译. 上海:机械工业出版社,2011.

[5] 林 丹. 4G 移动通信技术的现状与发展趋势探讨[J]. 科技信息,2013(24):241-241.

[6] Rossi M,Brunelli D. Forecasting data centers power consumption with the Holt-Winters method[C]//Proc of IEEE workshop on environmental,energy and structural monitoring systems. [s.l.]:IEEE,2015.

[7] 张小斐,田金方. 基于 ARIMA 模型的短时序预测模型研究与应用[J]. 统计教育,2006(10):7-9.

[8] 彭帅英,李广杰,彭 文,等. 基于改进遗传算法的 Holt-Winters 模型在采空沉陷预测中的应用[J]. 吉林大学学报:地球科学版,2013,43(2):515-520.

[9] 吴越强,吴文传,李 飞,等. 基于鲁棒 Holt-Winter 模型的超短期配变负荷预测方法[J]. 电网技术,2014,38(10):2810-2815.

[10] Li C,Chiang T W. Complexneurofuzzy ARIMA forecasting—a new approach using complex fuzzy sets[J]. IEEE Transactions on Fuzzy Systems,2013,21(3):567-584.

[11] 田 垅,刘宗田. 最小二乘法分段直线拟合[J]. 计算机科学,2012,39(6A):482-484.

[12] 段江娇. 基于模型的时间序列数据挖掘—聚类和预测相关问题研究[D]. 上海:复旦大学,2008.

[13] 微软中文. 大数据挖掘算法之:Microsoft 决策树算法[EB/OL]. [2014-10-13]. http://www.thebigdata.cn/JieJueFangAn/12096.html.

[14] Ziebarth N L,Abbott K C,Ives A R. Weak population regulation in ecological time series[J]. Ecology Letters,2010,13(1):21-31.

[15] Wu X,Zhu X,Wu G Q,et al. Datamining with big data[J]. IEEE Transactions on Knowledge & Data Engineering,2014,26(1):97-107.