



一种网络指标数据预测的方法

倪志刚

(中国移动通信集团福建有限公司 福州 350001)

摘要:随着通信行业规模的不断扩展,网管数据趋向多样化、复杂化和海量,其管理和优化工作更加困难,对数据指标的预测分析的要求也越来越高。如何有效、全面地预测网管数据,成为一个重要的课题。本文讨论了一种网络指标数据预测的方法。

关键词:网络指标;数据;预测

doi: 10.3969/j.issn.1000-0801.2013.03.024

A New Method of Network Index Data Forecast

Ni Zhigang

(China Mobile Group Fujian Co., Ltd., Fuzhou 350001, China)

Abstract: As the expansion of communication industry, the network data tends to be diverse, complex and numerous. Therefore, the management and optimization become more and more difficult. And the index data forecast requires higher and higher accuracy. How to manage the network data effectively and comprehensively turns to be a capital issue. A new method of network index data forecast was studied.

Key words: network index, data, forecast

1 引言

随着通信行业规模的不断扩展,网管数据趋向多样化、复杂化和海量,其管理和优化工作更加困难,对数据指标预测分析的要求也越来越高。怎样保证、提高数据质量,对网络指标数据质量进行预测,这无疑是摆在移动网络运营商面前的一大难题。

目前业界的数据预测分析普遍趋向以数据预测算法为核心,通过各类优秀的预测模型算法(如决策树算法、SVM 向量机、BP 神经网络算法、灰色预测算法等)达到数据预测的目标,这些模型算法在移动通信行业也被采用,在反复的研究试验使用中发现,模型算法虽为核心,但数据预测的结果取决于数据的样本和模型预测的使用方法。本文提出了一种数据预测的方法,以便有效地结合模型算法,提升数据预测的效果,这种网络数据预测方法的运用

将成为本文的阐述重点。

2 问题分析

针对网络数据预测的研究,引入了大量预测算法进行实践测试,在海量的实践中发现,数据预测模型的质量好坏与数据样本以及数据影响关联因素极其紧密。下文将讲述提出该预测方法的原因,介绍本方法如何针对业务,结合现有的模型算法,实现数据建模预测。

目前众多数据预测模型输出结果质量的差异,大部分原因是初期输入数据样本的差异,所以数据样本的准备极其重要,如果数据样本越全,数据量越大,那么预期的模型收敛速度变慢,运算资源消费也将扩大,但如果缩减数据样本又降低预测效果。如 BP 神经网络算法、决策树算法等预测模型,都有这个共同的特性。以下以 BP 神经网络算法为例,描述分析整体的模型结



构,如图1所示。

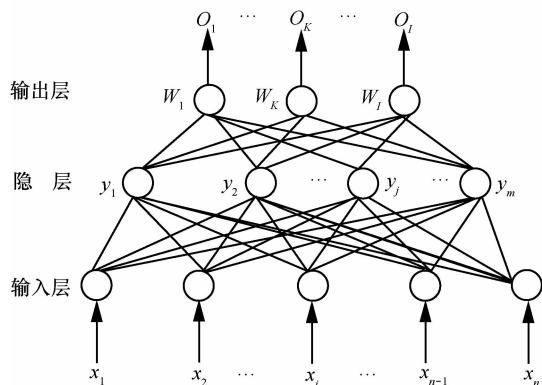


图1 BP神经网络算法的输入输出结构

在图1的模型中,输出层为 $o_k = f(\text{net}_k)$, $\text{net}_k = \sum_{j=0,m} w_{jk} y_j$, $k=1,2,\dots,m$; 隐层 $y_j = f(\text{net}_j)$, $\text{net}_j = \sum_{i=0,n} w_{ij} x_i$, $j=1,2,\dots,m$; 传递函数一般用 $f(x) = 1/(1+\exp(-x))$ (s型函数); 误差函数为 $Eh = 0.5 \times (\sum_{j=0,1} (t_{hj} - o_{hj})^2)$, t_{hj} 、 o_{hj} 分别为期望输出和网络的实际输出。以层概念划分,从输入层输入样本,经各隐层逐层处理后,传向输出层。若输出层的实际输出与期望输出不符合,则转入误差的反向传播阶段。误差反传是将输出误差以某种形式通过隐层向输入层逐层反传,并将误差分摊给各层的所有单元,从而获得各层单元的误差信号,此信号为修正各单元权值的依据。

通过以上介绍,可大致了解BP神经网络的使用,从数据角度来讲,该模型是一种局部搜索的优化方法,利用数据样本的训练建立神经网络,使得训练出的网络结构与期望值接近。但随着样本的过大化,也将会出现“过拟合”现象,原因是网络学习了过多的样本细节,学习出的模型已不能反映样本内含规律,所以如何把握学习度,是解决预测能力问题的关键。本文提出的预测方法中提到如何分析数据,梳理数据样本,从而达到对样本度的把控。

在本次的研究实践中发现,当把业务与算法模型结合起来进行实际的数据预测时,会出现众多的预测影响因素。如果不将影响因素考虑在内,显然会造成大量信息的丢失;如果将众多因素纳入关联,又会在预测方法上造成很大的麻烦,如基于一元N次回归预测、灰色预测GM(1,N)等预测模型中的N变量,反映了N-1个影响因素对某一变量的一阶导数的影响,但在实践中N影响因素的数量将决定实际预测的性能以及效果。

以灰色预测模型为例,如果考虑的模型由若干个相互影响的因素组成,设 $X_1^{(0)} = \{x_1^{(0)}(1), x_1^{(0)}(2), \dots, x_1^{(0)}(n)\}$ 为模型特征数据序列,则: $X_2^{(0)} = \{x_2^{(0)}(1), x_2^{(0)}(2), \dots, x_2^{(0)}(n)\}$, ...

$X_N^{(0)} = \{x_N^{(0)}(1), x_N^{(0)}(2), \dots, x_N^{(0)}(n)\}$ 为相关因素序列。

$X_i^{(1)}$ 为 $X_i^{(0)}$ 的1-AGO序列 ($i=1,2,\dots,N$), $Z_1^{(1)}$ 为 $X_1^{(1)}$ 的紧邻生成序列,则称 $x_1^{(0)}(k) + a z_1^{(1)}(k) = \sum_{i=2}^N b_i x_i^{(1)}(k)$ 为GM(1,N)灰色微分方程。

定义 $\hat{a} = [a, b_2, \dots, b_N]^T$ 为GM(1,N)灰色微分方程的参数列,根据最小二乘法可以得出 $\hat{a} = (B^T B)^{-1} B^T Y$, 称 $\frac{dx_1^{(1)}}{dt} + a x_1^{(1)} = b_2 x_2^{(1)} + b_3 x_3^{(1)} + \dots + b_N x_N^{(1)}$ 为GM(1,N)灰色微分方程的白化方程,也称影子方程:

$$\begin{aligned} x_1^{(1)}(t) &= e^{-at} \left[\sum_{i=2}^N \int b_i x_i^{(1)}(t) e^{at} dt + x_1^{(1)}(0) - \sum_{i=2}^N \int b_i x_i^{(1)}(0) dt \right] \\ &= e^{-at} [x_1^{(1)}(0) - t \sum_{i=2}^N b_i x_i^{(1)}(0) + \sum_{i=2}^N \int b_i x_i^{(1)}(t) e^{at} dt] \end{aligned}$$

通过该模型的算法分析发现,模型建立与影响因素紧密关联,但在实践中影响因素众多,将众多因素纳入模型中,对预测模型的运算资源消耗过重,因此如何从众多因素中筛选出主要的影响因素,达到预测关联降维的效果是本次需要解决的问题。

本文提出了一种数据指标预测的方法,利用主成分分析法进行分析降维,从众多的影响因素中选定主要的影响成分,从而解决如何在实际的数据预测中,筛选主要影响因素的问题,以支撑数据预测的稳定执行。

3 解决措施

结合现有的数据预测手段,依据中国移动的业务特性,根据移动网络业务的模型体系,提出了一种网络指标数据预测的方法,该方法主要分为4个步骤(“大点”),通过该4个步骤的细化,实现6个“小点”,实现数据分析建模预测的效果。以下将具体介绍该方法的主要实现思路,并利用SVM算法、BP神经网络算法、灰色预测算法3种算法进行建模预测,从而验证该方法的有效性。

3.1 数据预测具体步骤

有效的数据预测需要系统的分析。问题定义不清晰、数据质量不高、分析模型不正确,都会导致错误的预测结果。本文参考和结合数据挖掘流程,针对数据强化维度、强化因子分析,提出了一种网络数据预测的方法,该方法主要分成如下4个步骤,如图2所示。

3.1.1 数据分析

在数据预测的前期,要对数据进行分析处理,主要针

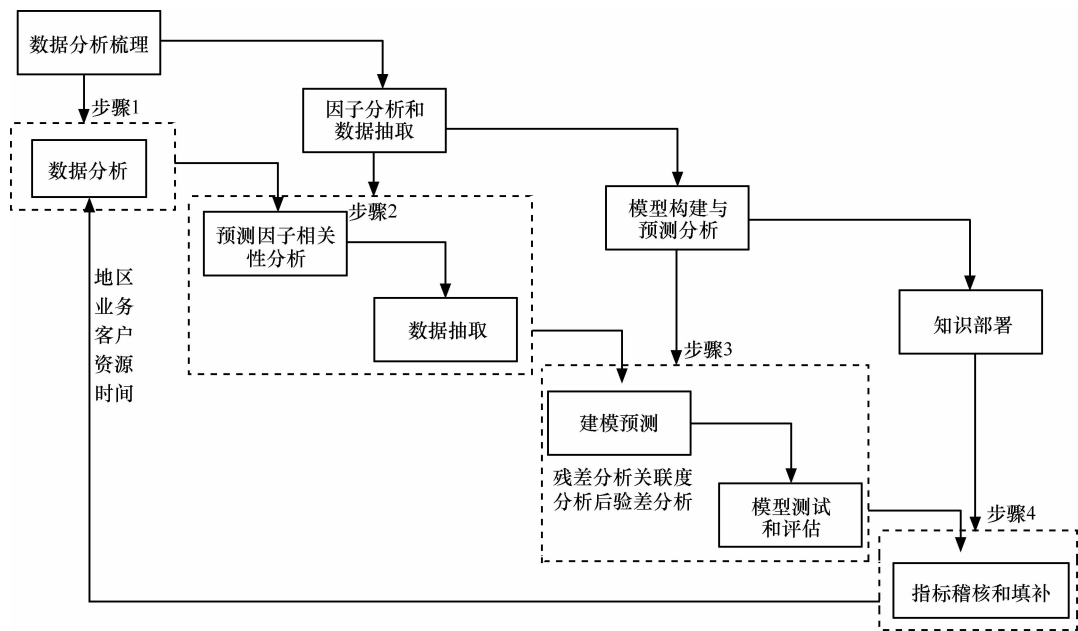


图2 数据预测方法的具体步骤

对预测对象的维度和影响因子进行分析梳理。影响关联因子是指与预测对象有关联关系的对象,根据笔者对移动网管网管业务的多年分析梳理,移动网管的数据特性有如表1所示的预测维度。

表1 网管数据特性的预测维度		
维度	说明	应用
时间维	不同的时间点 (时间段)	业务运行情况监控、趋势分析。在移动业务中,此维度为必选主因素之一
用户维	行业、集团客户	面向客户的分析与监控
区域维	地市、区域	面向区域的分析与监控
业务维	业务类别、业务	面向业务的分析与监控
资源维	所承载的基础网络	分析问题影响情况

根据以上维度,结合移动网络层的业务数据,例举如图3所示结构。

图3结构中,时间维度是网管数据的主要维度,是决定预测结果的主要因素,通过关联性分析法证实,其同周期时间维度的数据关联度最高,所以预测结果需要引入周期样本。如预测非节假日 GPRS 每日的忙时(18时)指标,即可利用非节假日每日忙时(18时)的历史数据作为数据样本。

通过以上维度分析,将数据进行梳理,划分出合理的数据样本,根据相应的维度,输入相应的样本,可有效提高预测的准备性、真实性。如福州移动 GPRS 流量,每日总流量较稳定,因此在进行 GPRS 每日的小时预测时,可以天为数据周期,福州为地市维度,利用每日同一小时时段数据

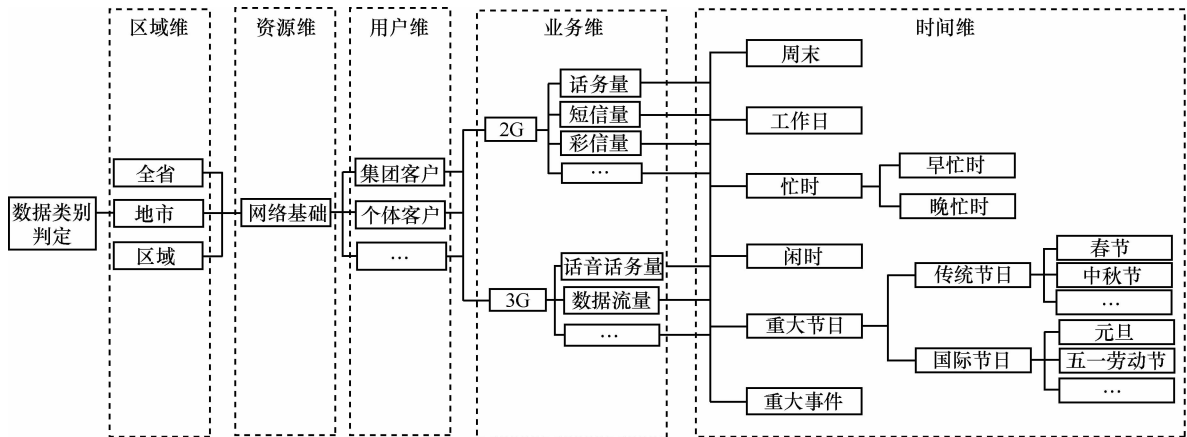


图3 移动网络层的业务数据结构



作为预测数据样本,完成数据预测建模,最终数据预测的效果较为明显。

3.1.2 预测因子相关性分析与数据抽取

同一预测对象有多个与之关联的预测因子,如果不加辨别地将这些因子都引入预测模型,势必无谓地增加模型的复杂度和分析计算的难度,因此本方案中使用主成分分析法对预测对象的预测因子进行分析筛选。

主成分分析法是一种降维的统计方法,借助于一个正交变换,将与其分量相关的原随机向量转化成与其分量不相关的新随机向量,这在代数上表现为将原随机向量的协方差阵变换成对角阵,在几何上表现为将原坐标系变换成新的正交坐标系,使之指向样本点散布最开的 p 个正交方向。然后对多维变量系统进行降维处理,使之能以一个较高的精度转换成低维变量系统,再通过构造适当的价值函数,进一步把低维系统转化成一维系统。

通过主成分分析法,可以有效剔除冗余的因子,降低计算的复杂性。最终通过相应的维度和关联因子抽取相应的样本数据。

3.1.3 建模预测与模型测试评估

当数据样本准备完成后,即可进行数据建模,数据建模预测的相应流程如图4所示。

根据图4对数据建模预测步骤进行相应描述。

- (1)数据引入:将已准备好的数据样本引入数据模型中。
- (2)预处理:对数据模型进行归一处理,提升模型换算

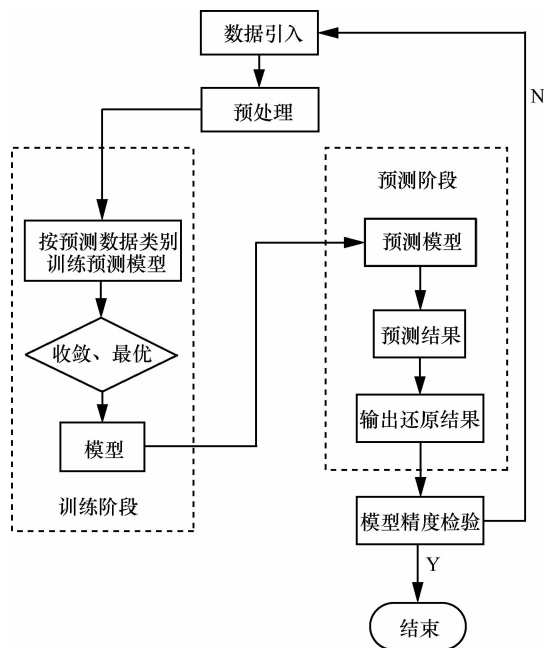


图4 数据建模预测的相应流程

效率。

(3)数据建模预测主要分为两个阶段:训练阶段和预测阶段,本文将利用3种算法,对归一后的数据模型进行数据预测训练,实现预测建模。训练阶段,利用预测算法对归一后的数据模型进行数据训练,检验该模型参数是否已经收敛或者达到最优,如果没有,则变换参数重新训练预测模型,直到得到最优的预测模型参数,即训练阶段结束;预测阶段,导入训练好的预测模型参数,结合样本数据进行预测,得到预测结果。

(4)预测模型构建完成后,采用精度检验方式进行模型校验,依次循环获得最准确的预测结果。精度校验的方式有残差验证、关联度验证、后验差验证。

- 残差验证:比较预测值和真实值之间的误差,是一种直观的描述方式。
- 关联度检验:主要反映预测数据对历史数据的拟合程度,通过查询经验表,可以估算模型建立是否合理。
- 后验差校验:主要表现误差的概率分布情况,以进一步论证模型的合理性。

3.1.4 指标稽核和填补

在数据部署环节,引用输出的预测结果与真实数据校验稽核,从中体现数据的准确性、波动性等,从而体现数据挖掘的效果,在现网中也可能会出现测量数据到达时延丢失的情况,即可利用预测结果实现数据填补的作用,体现数据的平滑性。

3.2 数据预测应用实例

结合移动业务应用场景,选取一类业务指标作为本次介绍的应用案例。在每个应用案例中,引用3种算法(SVM算法、灰色预测、BP神经网络)验证本次模型的可行性。

3.2.1 指标输入

以网管的IMS网络接通率(福建移动全省)为案例,利用此业务指标对模型进行数据实践。指标的相关算法如下:IMS网络接通率=主叫接通次数+被叫接通次数/主叫试呼次数+被叫试呼次数。

3.2.2 过程输出

(1)数据分析

针对全省IMS网络接通率指标,梳理出相关预测因子和指标维度,预测因子为:主叫接通次数、被叫接通次数、主叫试呼次数、被叫试呼次数。指标维度见表2。数据分析结果为获得数据样本。

表 2 指标维度

维度	说明
时间维	维度的类型为节假日、非节假日的繁忙时段、空闲时段、普通时段。忙时分分为早忙时 8~10 点、晚忙时 19~21 点,空闲时段时:0~7 点、普通时间段:11~23 点
区域维	全省
业务维	IMS 网络

(2)预测因子相关性

针对梳理出的 4 个预测因子进行主成分分析,计算得到主叫接通次数、被叫接通次数、主叫试呼、被叫试呼 4 个预测因子的贡献率,见表 3。

表 3 预测因子的贡献率

指标名称	主叫接通次数	被叫接通次数	主叫试呼次数	被叫试呼次数
贡献率	95.931 5%	2.951 3%	0.682 1%	0.435 1%

通过对以上数据的分析,可知主叫接通次数为预测因子主要成分。

当上述步骤完成后,建模初期的准备条件完成,以下步骤只需要根据应用算法进行数据建模预测。

3.2.3 数据预测及校验

利用 3 种算法进行数据预测,预测数据与真实数据的误差在专业人员制定的 5%范围内,证实了数据的有效性;证明提出的预测方法与预测算法的无相关性以及本文预测方法的可行性。通过梳理出的数据样本获得的 2012 年 8 月 27 日的预测校验结果介绍如下。

(1)SVM 预测校验

残差校验:SVM 预测残差校验结果见表 4。

关联度校验:3 个时段预测数据和真实数据的关联度分别为 0.950 2、0.868 9、0.923 9,取最大差的百分比为 50,即 $\rho=0.5$,根据经验公式,以上 3 个关联度均有效。

后验差检验:3 个时段预测数据的方差比和后验概率分别如下。

- 方差比为 0.431, $P=P\{|\Delta^{(0)}(i)-\bar{\Delta}^{(0)}|<0.674\ 5\ S_i\}=0.81$ 。
- 方差比为 0.423, $P=P\{|\Delta^{(0)}(i)-\bar{\Delta}^{(0)}|<0.674\ 5\ S_i\}=0.84$ 。
- 方差比为 0.476, $P=P\{|\Delta^{(0)}(i)-\bar{\Delta}^{(0)}|<0.674\ 5\ S_i\}=0.78$ 。

(2)BP 神经网络预测校验

残差校验:BP 神经网络预测残差校验结果见表 5。

关联度校验:3 个时段预测数据和真实数据的关联度分别为 0.816、0.712 2、0.824,取最大差百分比为 50,即 $\rho=0.5$,根据经验公式,以上 3 个关联度均有效。

后验差检验:3 个时段预测数据的方差比和后验概率

表 4 SVM 预测残差校验结果

时间	IMS 接通率 (实际值)	IMS 接通率 (预测值)	残差
2012/8/27 0:00	0.790 1	0.801 3	1.41%
2012/8/27 1:00	0.844 9	0.801 3	5.16%
2012/8/27 2:00	0.839 1	0.801 3	4.51%
2012/8/27 3:00	0.783 9	0.801 3	2.22%
2012/8/27 4:00	0.849 8	0.801 3	5.71%
2012/8/27 5:00	0.772 8	0.801 3	3.68%
2012/8/27 6:00	0.780 3	0.801 3	2.69%
2012/8/27 7:00	0.801 2	0.801 3	0.01%
2012/8/27 8:00	0.842 9	0.812 4	3.62%
2012/8/27 9:00	0.826 0	0.812 4	1.64%
2012/8/27 10:00	0.813 4	0.812 4	0.12%
2012/8/27 11:00	0.815 0	0.811 6	0.42%
2012/8/27 12:00	0.821 2	0.811 6	1.18%
2012/8/27 13:00	0.853 6	0.811 6	4.92%
2012/8/27 14:00	0.831 1	0.811 6	2.35%
2012/8/27 15:00	0.812 6	0.811 6	0.13%
2012/8/27 16:00	0.810 2	0.811 6	0.17%
2012/8/27 17:00	0.819 1	0.811 6	0.92%
2012/8/27 18:00	0.828 9	0.811 6	2.09%
2012/8/27 19:00	0.821 5	0.812 4	1.11%
2012/8/27 20:00	0.793 8	0.812 4	2.34%
2012/8/27 21:00	0.830 7	0.812 4	2.20%
2012/8/27 22:00	0.847 7	0.811 6	4.27%
2012/8/27 23:00	0.839 0	0.839 0	0

表 5 BP 神经网络预测残差校验结果

时间	IMS 接通率 (实际值)	IMS 接通率 (预测值)	残差
2012/8/27 0:00	0.790 1	0.819 1	3.67%
2012/8/27 1:00	0.844 9	0.815 3	3.50%
2012/8/27 2:00	0.839 1	0.806 2	3.93%
2012/8/27 3:00	0.783 9	0.771 2	1.62%
2012/8/27 4:00	0.849 8	0.778 2	8.42%
2012/8/27 5:00	0.772 8	0.784 9	1.57%
2012/8/27 6:00	0.780 3	0.815 7	4.54%
2012/8/27 7:00	0.801 2	0.820 1	2.36%
2012/8/27 8:00	0.842 9	0.813 8	3.46%
2012/8/27 9:00	0.826 0	0.813 8	1.48%
2012/8/27 10:00	0.813 4	0.813 8	0.05%
2012/8/27 11:00	0.815 0	0.818 2	0.40%
2012/8/27 12:00	0.821 2	0.813 5	0.94%
2012/8/27 13:00	0.833 6	0.811 3	2.67%
2012/8/27 14:00	0.831 1	0.818 2	1.55%
2012/8/27 15:00	0.812 6	0.818 2	0.69%
2012/8/27 16:00	0.810 2	0.818 2	0.99%
2012/8/27 17:00	0.819 1	0.818 2	0.11%
2012/8/27 18:00	0.828 9	0.816 9	1.45%
2012/8/27 19:00	0.821 5	0.813 8	0.94%
2012/8/27 20:00	0.793 8	0.813 8	2.51%
2012/8/27 21:00	0.830 7	0.813 8	2.04%
2012/8/27 22:00	0.847 7	0.804 6	5.09%
2012/8/27 23:00	0.839 0	0.790 6	5.77%



分别如下。

- 方差比为 0.492, $P=P\{|\Delta^{(0)}(i)-\bar{\Delta}^{(0)}|<0.674\ 5\ S_i\}=0.77$ 。
- 方差比为 0.423, $P=P\{|\Delta^{(0)}(i)-\bar{\Delta}^{(0)}|<0.674\ 5\ S_i\}=0.74$ 。
- 方差比为 0.398, $P=P\{|\Delta^{(0)}(i)-\bar{\Delta}^{(0)}|<0.674\ 5\ S_i\}=0.81$ 。

(3) 灰色预测校验

残差校验:灰色预测残差校验结果见表 6。

表 6 灰色预测残差校验结果

时间	IMS 接通率 (实际值)	IMS 接通率 (预测值)	残差
2012/8/27 0:00	0.790 1	0.790 1	0
2012/8/27 1:00	0.844 9	0.816 7	3.33%
2012/8/27 2:00	0.839 1	0.816 1	2.75%
2012/8/27 3:00	0.783 9	0.815 5	4.03%
2012/8/27 4:00	0.849 8	0.814 8	4.12%
2012/8/27 5:00	0.772 8	0.814 2	5.36%
2012/8/27 6:00	0.780 3	0.813 5	4.26%
2012/8/27 7:00	0.801 2	0.812 9	1.46%
2012/8/27 8:00	0.842 9	0.842 9	0
2012/8/27 9:00	0.826 0	0.819 1	0.84%
2012/8/27 10:00	0.813 4	0.818 1	0.58%
2012/8/27 11:00	0.815 0	0.822 3	0.90%
2012/8/27 12:00	0.821 2	0.823 4	0.26%
2012/8/27 13:00	0.853 6	0.824 4	3.42%
2012/8/27 14:00	0.831 1	0.825 5	0.67%
2012/8/27 15:00	0.812 6	0.826 6	1.72%
2012/8/27 16:00	0.810 2	0.827 7	2.16%
2012/8/27 17:00	0.819 1	0.828 7	1.17%
2012/8/27 18:00	0.828 9	0.829 8	0.11%
2012/8/27 19:00	0.821 5	0.817 1	0.54%
2012/8/27 20:00	0.793 8	0.816 1	2.80%
2012/8/27 21:00	0.830 7	0.815 0	1.89%
2012/8/27 22:00	0.847 7	0.830 9	1.99%
2012/8/27 23:00	0.839 0	0.839 0	0

关联度校验:3 个时段预测数据和真实数据的关联度分别为:0.784、0.813、0.845, 取最大差的百分比为 50, 即 $\rho=0.5$, 根据经验公式, 以上 3 个关联度均有效。

后验差检验:3 个时段预测数据的方差比和后验概率分别如下。

- 方差比为 0.424, $P=P\{|\Delta^{(0)}(i)-\bar{\Delta}^{(0)}|<0.674\ 5\ S_i\}=0.84$ 。
- 方差比为 0.437, $P=P\{|\Delta^{(0)}(i)-\bar{\Delta}^{(0)}|<0.674\ 5\ S_i\}=0.77$ 。
- 方差比为 0.345, $P=P\{|\Delta^{(0)}(i)-\bar{\Delta}^{(0)}|<0.674\ 5\ S_i\}=0.83$ 。

4 结束语

本次提出的针对移动网络指标数据的预测方法, 对影响数据质量的因子、维度等一些特别因素与实际的移动网管业务参数进行数据分析预测, 弥补了现有预测方法的不足, 具有以下优点:

- 结合实际的移动网管业务参数, 对业务数据指标进行预测, 提升了业务的预见性、可控性;
- 能够结合业务场景, 针对节假日等一些特别因素进行数据预测分析;
- 通过该方法的思路指导, 有效地分析样本、理解数据, 结合已有的模型算法构建合理的数据模型, 提升网络指标数据的预测效果。

参考文献

- 1 Price J. Oracle Database 11g SQL. McGraw-Hill Osborne Media, 2007
- 2 Feuerstein S, Priby B. Oracle PL/SQL Programming: Covers Versions Through Oracle Database 11g Release 2 (Animal Guide). O'Reilly Media, 2009
- 3 邓聚龙. 灰色预测与决策. 武汉:华中理工大学出版社, 1986
- 4 石山铭. 神经网络预测原理与应用. 系统工程学报, 1994(1)
- 5 陈蓉. 话务量分析和多种预测模型的比较研究. 北京邮电大学硕士学位论文, 2008
- 6 Cristianini Nello, Shawe-Taylor J 著. 李国正译. 支持向量机导论(第 1 版). 北京:电子工业出版社, 2004
- 7 邓乃杨, 田英杰. 支持向量机—理论、算法与拓展(第 1 版). 北京:科学出版社, 2009
- 8 陈颖. SAS 数据分析系统教程 (第 1 版). 上海: 复旦大学出版社, 2008
- 9 熊春光. 科学与工程计算方法——数学. 北京: 北京理工大学出版社, 2011
- 10 Simon Haykin 著. 叶世伟, 史忠植译. 神经网络原理. 北京: 机械工业出版社, 2004

[作者简介]



倪志刚, 男, 学士, 中国移动通信集团福建有限公司网管支撑室副室经理、工程师, 中国移动通信集团 CMOSS 规划组成员, 负责中国移动通信集团福建有限公司网管 OSS 规划项目, 主要研究方向为网管规划、云计算。

(收稿日期: 2013-03-08)