呼叫中心分块回归话务量预测

张沪寅,胡瑞芸,何 政
ZHANG Huyin, HU Ruiyun, HE Zheng

武汉大学 计算机学院,武汉 430072
Computer School, Wuhan University, Wuhan 430072, China

ZHANG Huyin, HU Ruiyun, HE Zheng. Block regression traffic prediction model for call center. Computer Engineering and Applications, 2016, 52(12):90-94.

Abstract: In order to obtain the prospective traffic data, solve the seats arrangement problem of call center, realize the rational allocation of human resources, block regression traffic prediction model, based on support vector machine and *K*-nearest neighbor algorithm is proposed(SKBR), after analyzing the characteristics of historical traffic data. According to the date type, traffic can be divided into weekday traffic, weekend traffic and holiday traffic, and different model is used to predict the corresponding traffic. Taking the traffic of a province electric power call center for example, experiments are carried on the MATLAB platform. Results show that compared with the SVM model and improved SVM model for its method of searching parameters, SKBR model has improved the prediction accuracy.

Key words: traffic; prediction; support vector machine; nearest neighbor algorithm; prediction accuracy

摘 要: 为获得前瞻性话务量数据,解决呼叫中心坐席安排的问题,实现人力资源合理配置,分析历史话务量特性,提出了基于支持向量机和 *K* 近邻算法的分块回归(SKBR)话务量预测模型。将话务量按日期类型分为工作日话务量、周末话务量以及节假日话务量,采用不同的模型预测相应的话务量。以某省电力呼叫中心话务量为例,在Matlab平台上进行实验。结果证明,相比SVM模型和改进寻参方法的SVM模型,SKBR模型在预测准确性上有所提升。

关键词: 话务量; 预测; 支持向量机; 近邻算法; 预测准确性
文献标志码: A **中图分类号:** TP39 doi:10.3778/j.issn.1002-8331.1501-0160

1 引言

随着人们对电力服务质量的要求不断提高,国家电网已经把提供优质的供电服务提升到了发展战略的新高度。95598呼叫中心作为供电企业与用电客户的交流平台,实现了24 h不间断随时提供服务的可能。话务量是呼叫中心进行客服坐席安排的依据,针对不同的话务量安排相应的坐席,才能实现保证呼叫中心服务质量和实现人力资源最优配置的双重目标[1]。传统的排班模式,需要经验丰富的排班师对话务量进行估计,工作量较大,且包含一定的人为的主观因素,无法满足实际生产需求。因此,如何对话务量进行科学准确的预测已经成为了一个亟待解决的问题。

话务量是一种随机的、动态的时间序列[2],受天气、季节、节假日等因素的影响,呈现复杂的变化趋势。目前,已有一些预测工具被应用于各种通信系统的话务量预测中,比如,ARIMA模型[3]、多元线性回归[4]、Kalman滤波[5]、BP神经网络[6]等,并都取得了一定的成果。但ARIMA模型要求序列是平稳序列,多元线性回归和Kalman滤波模型相对简单,难以满足话务量的复杂变化,BP神经网络需要利用大量样本训练模型且训练速度较慢[6],不满足现阶段对话务量预测的要求。针对以上不足以及话务量自身的特点,提出了一种基于支持向量机(Support Vector Machine,SVM)和 *K* 近邻算法(*K*-Nearest Neighbor,*K*NN)的分块回归预测模型(SKBR)。

基金项目: 高等学校博士学科点专项科研基金(No.20130141110022);武汉市科学技术局(No.201302038)。
作者简介: 张沪寅(1962—),男,博士研究生,教授,研究领域为数据挖掘,网络QoS,新一代网络体系结构;胡瑞芸(1991—),女,硕士研究生,研究领域为数据挖掘,机器学习,E-mail:ryh_ok@whu.edu.cn;何政(1976—),男,博士后,讲师,研究领域为数据挖掘,机器学习,优化算法。
收稿日期: 2015-01-13 **修回日期:** 2015-03-31 **文章编号:** 1002-8331(2016)12-0090-05
CNKI **网络优先出版:** 2015-07-03, http://www.cnki.net/kcms/detail/11.2127.TP.20150703.1610.014.html

针对不同类型日期定制不同的预测模型,实验证明,话务量的预测准确性有所提高。

2　分块回归预测模型

分块回归预测模型的主要思想是不同类型日期使用不同的模型进行预测。话务量影响因素复杂,故特征维较多,工作日话务量相对周末话务量数据量要大得多,因此在使用支持向量机回归模型进行预测之前先运用聚类算法对训练集数据进行筛选,以缩短建模时间,而节假日话务量数据量少,时序间隔较长,随机性大,无明显规律性,因此采用近邻算法进行预测。模型整体框图如图1所示。

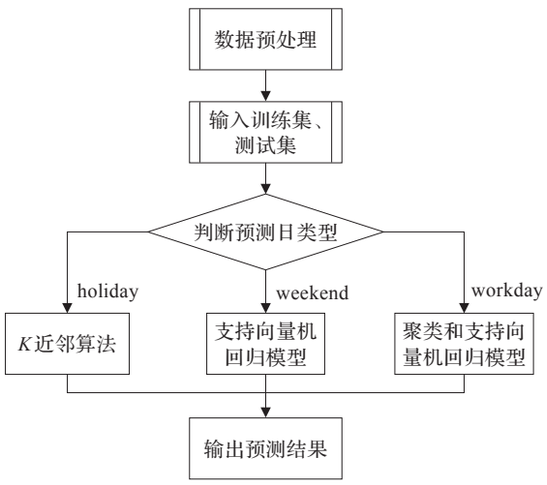

图1　模型整体框图

2.1　支持向量机

支持向量机基于结构风险最小化理论[7],针对线性可分情况进行分析,对于线性不可分的情况通过非线性映射转化为在高维特征空间中线性可分的,在解决小样本、非线性和高维模式识别问题中表现出许多特有的优势[8]。训练模型时采用网格搜索算法[9]与交叉验证[10]相结合的方法确定模型参数,并用得到的最优模型进行预测。然而运用目前常用的网格搜索算法求惩罚参数 C 和核函数参数 γ 时,所采用的搜索路线形式为 2^x,随着 x 值的增加,搜索间距逐渐增大,可能出现漏过最优参数的情况,因此采用一种新的寻参表达式,形式如下:

$$C=ax^2+bx+c \tag{1}$$

$$\gamma=ae^{bx} \tag{2}$$

因支持向量机模型对 γ 的取值尤为敏感,且其取值范围一般为0~1,所以公式(2)中 x 的取值小于零且变化步长小。以下比较了上述两种表达式形式的变化趋势。

图2中左侧带星形标记的曲线是搜索参数 γ 最优值的路线,n 的取值范围是 -15~0,上升步长为0.5,右侧带圆形标记的曲线是搜索参数 C 最优值的路线,m 的取值范围是0~8,上升步长为0.8。可以看出,蓝色曲线随着 m 或 n 值的增加上升速度越来越快,相邻两点间的间距逐渐增大,而红色曲线上升平缓,相邻两点间的间距较均匀,这对逐步搜索参数最优值是有利的。

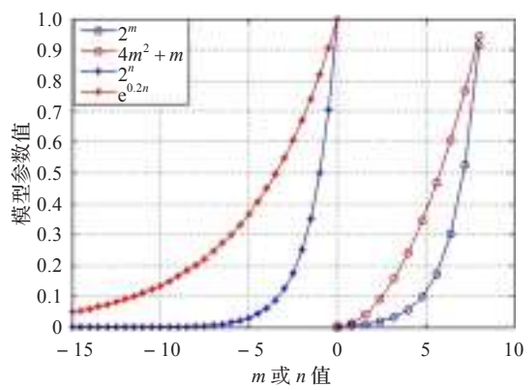

图2　搜索SVM模型最优参数的路线比较

2.2　模糊C均值聚类

根据支持向量机理论,SVM回归的决策函数由样本中的支持向量(Support Vector,SV)决定[11],然而在优化学习的过程中将耗费较多的时间在非SV上[12],因此在训练之前对非SV进行去除,对减少计算量、缩短训练时间是必要的。为解决这个问题,本文中采用模糊C均值聚类的方法来对历史数据进行筛选,选择与预测日最相似的数据作为训练集进行学习。

2.3　*K*近邻算法

*K*NN算法是著名的模式识别统计学方法,基本思想是"近朱者赤,近墨者黑"[13],由邻居类型来推断其所属的类别。

*K*NN算法一般用于分类,这里用来做预测,只需将原算法求分类的步骤改为由 k 个邻居的话务量值决策预测日的话务量。k 个邻居与预测样本的距离大小不一,存在"近大远小"的关系,因此不采用通常取邻居的均值作为预测值的方法,而是将与邻居间的距离作为权重系数代入计算。预测值表达式为:

$$y=\frac{\sum_{i=1}^{k}y_i\times\frac{1}{d_i}}{\sum_{i=1}^{k}\frac{1}{d_i}} \tag{3}$$

其中,k 表示邻居个数,$y_i, i=1,2,\cdots,k$ 为邻居话务量,d_i 为对应的距离。

3　呼叫中心话务量预测

3.1　话务量影响因素

电力呼叫中心担任着故障报修、业务咨询、投诉等职责,这些业务直接影响话务量的多少,而这些业务本身受多方面因素的影响。

(1)天气:恶劣天气情况下,如雷雨、大风、降雪、高温

等,易导致供电设备损坏或供电不足而出现用电故障。

(2)季节:每年的7月至8月是迎峰度夏期,12月至2月是枯水期[14],在这两段季节期间由于酷暑、用电量大或者寒冷、水力发电能源不足造成停电的可能性较大。

(3)节假日:节假日与工作日对用电的需求不同,话务量自然不同。

(4)供电量:当电力公司对区域的供电量出现不足时,相应区域的话务量就会增多。

(5)供电设施、资费政策:陈旧老化的供电设施,在恶劣天气下更易损坏,因此设施的质量、新旧程度等对话务量也有一定影响。当电力公司出台新的用电政策、资费标准,将会有较多用户进行信息咨询,话务量也会随之变化。

影响话务量的因素众多且关系复杂。如上述提及的天气和季节是相关的,一般来说,夏季的温度与降雨普遍高于冬季,在高温、大风等恶劣天气下供电设施更易损坏。因此,在进行话务量预测时,需要合理选择这些影响因素。

3.2　话务量特性分析

图3显示了某省2013年5月和2014年5月的日话务量数据,其中方形标识是节假日,星形标识是工作日,圆形标识是周末。由图可知,工作日话务量普遍高于周末和节假日话务量,节假日话务量略高于周末。从图中还可以看出,话务量具有一定的年周期性,主要表现在节假日话务量上,不同年份的同一节假日,话务量是相似的。因此,在预测时应将话务量按日期类型分为工作日、周末、节假日三类,分别选择相应的历史数据进行预测。

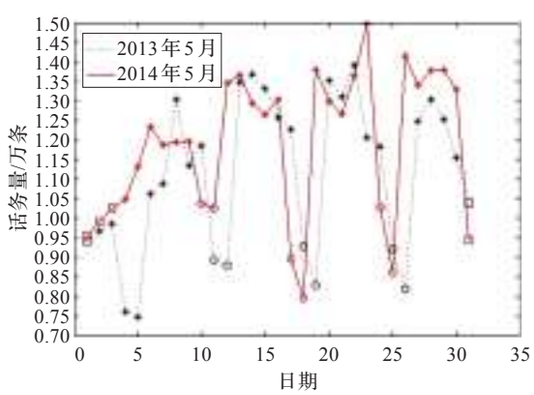

图3　2013年5月与2014年5月话务量对比图

图4显示了2013年1月至2014年6月期间的月话务量均值,分析发现话务量具有年周期性,随季节变化较明显。最显著的是夏季7月和8月,相比其他季节话务量激增,这与前文分析的话务量受天气、季节的影响是相吻合的,国家电网将这两个月视为迎峰度夏期,故此时间段内的话务量不与其他月份采用相同的方法进行预测,本文不予讨论。其他月份的话务量均值分布相对波动较小,但仍然存在一定差异。

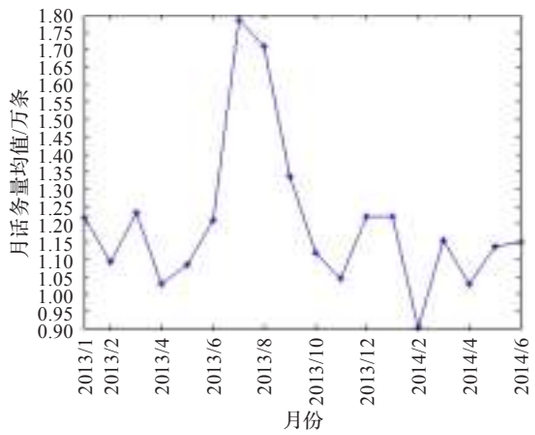

图4　月话务量均值分布图

使用Matlab对话务量以月为单位进行分布检验,发现大部分月份的话务量服从置信度为0.95的正态分布,记为 $X\sim N(\mu,\sigma^2)$,其中 X 为月话务量序列,μ 为序列均值,σ 为标准差。这一特性是话务量预处理时采取以月为单位进行的依据。

3.3　数据预处理

数据在收集整理的过程当中难免会出现问题或者错误,在进行预测之前需要先对数据的准确性和有效性进行识别和处理。经过话务量特性分析看出每月话务量差别较大,因此以月为单位进行检验。

3.3.1　异常数据识别

(1)传统方法:以月平均值和2倍标准差的范围($[\mu-2\times\sigma,\mu+2\times\sigma]$)为判断依据,落入区间范围内的数据为正常数据,超出区间范围的数据为异常数据。

(2)计算样本值概率密度:对满足正态分布的月话务量序列,计算样本值的概率密度,设定概率密度阈值,认为概率低于这个阈值的事件为小概率事件,即取到这个样本值是小概率事件,将这样的样本值判断为异常数据,反之为正常数据。

考虑到在月份数据中可能存在某些过高或过低的数据对当月的平均值和标准差产生较大的影响,而使月平均值不能准确反映当月话务量的整体水平,故在满足正态分布的月份,将这两种方法结合使用,以对异常数据进行更全面、准确的判断。对不满足正态分布的月份,采用方法(1)进行识别。

3.3.2　异常数据处理

识别出异常数据后,进行修复的基本思想是用相邻的同类型日话务量均值代替。具体方法是,若异常日是工作日,则用当天前后各三天的工作日正常话务量均值代替;若异常日是周末,则用相邻的前后各一周以及本周的周末正常话务量均值代替;若异常日是节假日,则用前一年同一假期以及当天前后各一天的正常话务量均值代替。

3.4　建模与预测

预测模型的输入与输出:模型的输入为对预测日话务量影响较大的历史话务量以及影响话务量的主要因素组成,输出为预测日的话务量预测值。依据文献[10]在预测电力负荷时利用灰色关联分析的方法确定模型的输入为:预测日的前一天、前两天、前三天、前一周话务量,月份 M ,季度 S ,星期 W ,最低温 MT ,平均温 AT ,雨 R ,雪 SW ,风 WD 。假设 $P(d)$ 为话务量预测值, $A(d)$ 为话务量实际值, d 为预测日,则话务量预测表达式可表示为:

$$P(d)=f(T(d-1),T(d-2),T(d-3),T(d-7),M(d),S(d),W(d),MT(d),AT(d),R(d),SW(d),WD(d)) \quad (4)$$

其中,月份、季度等离散变量通过数值量化表示,比如,月份可取的值为1~12之间的整数。

4　实验结果与分析

借助Matlab平台,libsvm工具箱,依据所建立的预测模型,以2013年1月至2014年4月的数据为训练集,预测2014年5月1日至5月15日的话务量。

实际值和预测值之间的差异,即预测误差,可以反映预测的准确性。预测误差越小,准确性越高,反之预测误差越大,准确性越低。可采用常用的平均绝对相对误差($MAPE$)[15]来衡量:

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\frac{|y'_i-y_i|}{y_i} \quad (5)$$

其中, y_i 为实际值, y'_i 为预测值。

使用SKBR模型进行预测的结果如表1。

表1　SKBR模型预测结果

日期	日期类型	实际值	预测值	预测误差
5月1日	3	9 572	8 185.2	0.145
5月2日	3	9 912	8 806.5	0.112
5月3日	3	10 252	9 257.4	0.097
5月4日	1	10 491	10 815.7	0.031
5月5日	1	11 331	12 274.0	0.083
5月6日	1	12 349	12 167.0	0.015
5月7日	1	11 887	12 064.7	0.015
5月8日	1	11 961	12 012.5	0.004
5月9日	1	11 973	12 279.1	0.026
5月10日	2	10 380	9 467.1	0.088
5月11日	2	10 250	9 811.1	0.043
5月12日	1	13 459	12 846.7	0.045
5月13日	1	13 678	12 969.2	0.052
5月14日	1	12 944	12 776.2	0.013
5月15日	1	12 651	12 445.1	0.016

计算得出SKBR模型的 $MAPE$ 值为0.052,误差较小,准确性较高。对同一数据集,采用单一SVM模型以及改进了寻参方法的SVM模型(ISVM, Improved SVM)进行预测,三种模型预测结果与实际值的拟合情况如图5。

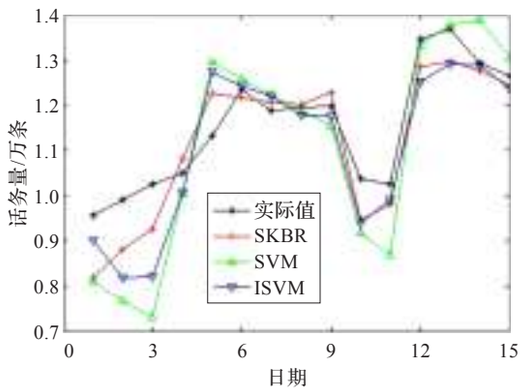

图5　三种模型的预测值与实际值的拟合图

SKBR模型整体上反映了话务量的变化趋势,但节假日预测值与实际值之间的差距仍然较大,ISVM模型预测值相对SVM模型更好地拟合了实际值,说明改进网格搜索算法的搜索路径有利于搜索到模型的最优参数。表2列出了各个模型的总体预测误差,以及不同日期类型的预测误差,可以看出SKBR模型的总体误差相对SVM模型和ISVM模型分别提高了3%和1%,工作日预测误差均提高了1%,周末预测误差相对SVM模型提高了7%,节假日预测误差分别提高了10%和3%。但SKBR模型的节假日预测误差仍然较大,这与话务量中节假日类型数据量小有一定关系,也可能与KNN算法中找邻居时使用普通的欧式距离进行度量有关,因为这样忽略了话务量影响因素对话务量作用程度不同的事实。

表2　三种模型预测误差对比

预测误差	总体	工作日	周末	节假日
SVM 模型	0.089	0.040	0.134	0.222
ISVM模型	0.063	0.038	0.063	0.144
SKBR模型	0.052	0.030	0.065	0.118

此外,若训练集未经过异常识别或仅使用传统方法进行异常识别,运用以上三种模型对测试集进行预测得到的总体平均预测误差如表3。

表3　不同异常处理方法的预测误差对比

异常处理方法	未处理	传统方法
SVM模型	0.112	0.098
ISVM模型	0.096	0.083
SKBR模型	0.080	0.068

结合表2和表3的实验结果可以看出,不论采用哪种预测模型进行预测,对训练集采取的异常识别方法直接影响预测误差的大小。传统方法以及结合计算样本值概率密度的方法都在一定程度上有利于提高预测准确性,两种方法结合使用的表现则更优,说明通过计算样本值概率密度判断样本值是否异常是可行的。

5　结束语

本文对电力呼叫中心话务量进行了特性分析、特征

提取与量化,并在数据预处理的过程中提出了采用计算样本值概率密度与传统方法相结合进行异常识别的方法,避免了传统方法的不足,最后按照话务量特性在改进网格搜索算法的基础上建立了SKBR模型,通过对工作日话务量、周末话务量和节假日话务量进行分块建模,采用不同的模型预测相应的话务量。同时,应用SVM模型和改进的SVM模型对相同的数据集进行预测实验,并与SKBR模型的实验结果相对比。实验证明SKBR模型能较好地拟合实际值,总体上提高了预测准确性,但节假日话务量预测效果还有待提高,需要进一步的研究。

参考文献:

[1] 牟颖,王俊峰,谢传柳,等.大型呼叫中心话务量预测[J].计算机工程与设计,2010,31(21):4686-4689.

[2] Moungnoul P,Laipat N,Hung T T,et al.GSM traffic forecast by combining forecasting technique[C]//Fifth International Conference on Information,Communications and Signal Processing,2005:429-433.

[3] Li Shulan,Huang Hongqiong,Zhu Daqi,et al.The application of space-time ARIMA model on traffic flow forecasting[C]//International Conference on Machine Learning and Cybernetics,2009:3408-3412.

[4] 王勇,黄国兴,彭道刚.带反馈的多元线性回归在电力负荷预测中的应用[J].计算机应用与软件,2008,25(1):82-84.

[5] 龙通彬.基于卡尔曼滤波的呼叫中心话务量预测[J].计算机工程与设计,2013,34(12):4405-4409.

[6] 邓波,李健,孙涛,等.基于神经网络的话务量预测[J].成都信息工程学院学报,2008,23(5):518-521.

[7] Han Rui,Jia Zhenghong,Qin Xizhong,et al.Application of support vector machine to mobile communications in telephone traffic load of monthly busy hour prediction[C]//International Conference on Natural Computation,2009:349-353.

[8] 丁世飞,齐丙娟,谭红艳.支持向量机理论与算法研究综述[J].电子科技大学学报,2011,40(1):2-10.

[9] 刘道文,忽海娜.基于网格搜索支持向量机的网络流量预测[J].计算机应用与软件,2012,29(11):185-187.

[10] 关颖.支持向量机在电力系统短期负荷预测中的应用[D].天津:天津大学,2006.

[11] Theja P V V K,Vanajakshi L.Short term prediction of traffic parameters using support vector machines technique[C]//International Conference on Emerging Trends in Engineering and Technology,2010:70-75.

[12] 陈电波,徐福仓,吴敏.基于聚类和支持向量机的话务量预测模型[J].控制工程,2009,16(2):195-198.

[13] 吴波,朱昌杰,任逸卿.文本分类技术探究[J].宿州学院学报,2012,27(5):19-23.

[14] 艾勇.电力呼叫中心话务量的指数平滑预测方法[J].中南民族大学学报:自然科学版,2012,31(3):81-84.

[15] 刘童,孙吉贵,张永刚.用周期模型和近邻算法预测话务量时间序列[J].吉林大学学报:信息科学版,2007,25(3):239-245.

(上接11页)

[5] Medina Ayala A I,Andersson S B,Belta C.Temporal logic control in dynamic environments with probabilistic satisfaction guarantees[C]//2011 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS),2011:3108-3113.

[6] Lahijanian M,Andersson S B,Belta C.Temporal logic motion planning and control with probabilistic satisfaction guarantees[J].IEEE Transactions on Robotics,2012,28(2):396-409.

[7] Ding Xuchu,Pinto A,Surana A.Strategic planning under uncertainties via constrained Markov decision processes[C]//2013 IEEE International Conference on Robotics and Automation(ICRA),2013:4568-4575.

[8] Baier C,Katoen J P.Principles of model checking[M].Cambridge:MIT Press,2008.

[9] Kwiatkowska M,Norman G,Parker D.Advances and challenges of probabilistic model checking[C]//Proc 48th Annual Allerton Conference on Communication,Control and Computing,2010:1691-1698.

[10] Parker D A.Implementation of symbolic model checking for probabilistic systems[D].University of Birmingham,2002.

[11] 王立权,孟庆鑫,郭黎滨,等.护士助手机器人的研究[J].中国医疗器械信息,2003,9(4):21-23.

[12] 张健,张国山,邵志刚,等.面向医院的移动机器人导航系统设计[J].微计算机信息,2007,23(20):200-202.

[13] Hinton A,Kwiatkowska M,Norman G,et al.PRISM:a tool for automatic verification of probabilistic systems[C]//12th International Conference on Tools and Algorithms for the Construction and Analysis of Systems(TACAS'06),2006:441-444.

[14] Kwiatkowska M,Norman G,Parker D.Probabilistic symbolic model checking with PRISM:a hybrid approach[J].International Journal on Software Tools for Technology Transfer(STTT),2004,6(2):128-142.

[15] PRISM-probabilistic symbolic model checker[EB/OL].[2014-05-10].http://www.prismmodelchecker.org/.