基于用户行为模型的客流量分析与预测①

程求江 [1,2], 彭艳兵 [2]

[1](武汉邮电科学研究院 光纤通信技术与网络国家重点实验室, 武汉 430074)
[2](南京烽火星空通信发展有限公司 研发部, 南京 210019)

摘 要: 为了预测无线城市接入中商圈的短时客流量, 通过分析顾客购物行为模式, 提出了一种基于停留时间和区间活跃度的身份识别方案, 用于区分工作人员和顾客; 采用二元线性回归方法对停留时间和活跃次数进行置信水平为 95%的拟合, 分析了不同拟合参数对预测的影响. 实验结果表明: 停留时间和活跃度用于区分身份信息合理有效, 且在时间阈值为 3 小时, 活跃度阈值为 2 次时, 用小波神经网络预测效果最好.

关键词: 停留时间; 活跃度; 身份识别; 小波神经网络; 预测

Customer Traffic Analysis and Forecast Based on User's Behavior Model

CHENG Qiu-Jiang[1,2], PENG Yan-Bing[2]

[1](State Key Laboratory of Optical Communication Technologies and Networks, Wuhan Research Institute of Posts and Telecommunications, Wuhan 430074, China)
[2](Research and Development Department, Nanjing FiberHome Star Communication Development Co.Ltd., Nanjing 210019, China)

Abstract: In order to forecast the short-term customer flow in trading area under wireless access, through analyzing of customer shopping behaviors, this paper presents an identification scheme based on staying time and activeness to distinguish the staff and customers. We use the binary linear regression method to fit the data under confidence level of 95%, and analyze the influence of different parameters to predict. Experimental results show that the staying time and activeness are reasonable and effective to distinguish the identity information, when time threshold is 3 and activeness threshold is 2, the wavelet neural network prediction effect is best.

Key words: staying time; activeness; identity recognition; wavelet neural network; forecasting

移动互联网的发展和智能终端的普及, 带来了流量风暴问题, 为了减轻蜂窝网络负载, 近年来提出了无线城市建设, 目前已在各大中型城市如火如荼的进行. 设备接入到 Wi-Fi 热点时, 服务端会产生一条日志信息, 挖掘海量日志信息中隐藏的规律, 能够有效指导商业行为.

客流量是商业分析中重要考虑因素, 特别是短时客流量, 其对商业抉择、人群导向、后勤保障有重要意义. 短时预测问题具有高度的非线性和不确定性, 并且同时间相关性较强, 该类问题常用的方法主要有以自回归移动均值为代表的回归模型预测方法[1]和以神经网络为代表的机器学习方法[2-5], 前者在预测精度和刻画复杂序列的全部特征方面劣于神经网络[2,3]. 小波神经网络是一种优化的 BP 神经网络, 它继承了小波分析的多分辨率分析和处理能力又保留了神经网络在函数逼近上具有自学习、自适应、容错等优点, 且克服了 BP 网络常不收敛, 并且收敛速度慢的问题, 在预测精度方面有了较大提高[4,5].

为了预测客流量, 需要对日志信息中每个 Mac 地址所代表的用户行为进行分析, 以区分场所工作人员和顾客. 传统的行为分析主要基于 Web 数据挖掘[6,7], 通过分析用户上网点击行为, 浏览内容, 挖掘用户的行为特征, 然后进行相关推荐活动, 达到精准营销的目的. 随后出现了微信、微博为代表的基于地理位置

① 基金项目:江苏省科技支撑计划(BE2011173)
收稿时间: 2014-07-10;收到修改稿时间:2014-08-18

的移动社交网络[8,9], 用户通过签到、发微博、发说说等行为分享周边信息, 数据的稀疏性问题, 影响了对用户的行为特征的挖掘, 推荐效果不好. 然而 Mac 地址数据获取简单, 数据量大, 只要用户开着 Wi-Fi 经过无线 AP, 该 AP 便会记录下设备的 Mac 地址, 并且每个设备的 Mac 地址信息是唯一的, 数据真实可靠, 非常适合商业分析.

一条日志信息包含三个元素: AP 的 Mac 地址、用户设备的 Mac 地址、时间信息, 而停留时间和活跃度能很好的刻画一个人购物时的行为模式, 本文首先对工作人员和顾客两种身份进行识别, 然后对顾客的短时流量进行预测.

1 模型建立

1.1 身份识别模型

Wi-Fi 热点的扫描周期为 0.1 秒, 理论上每隔 0.1 秒, AP 会产生一条日志信息, 并将信息发送给汇聚中心, 为了减轻汇聚中心服务器压力, 日志信息生成策略为: 不同 AP 相互之间互不影响, 单一 AP 若在某时间点扫描到用户 A, 并发送了该条记录, 则单位时间粒度内扫描到用户 A 的信息会合并, 只保留最早的记录. 停留时间的计算方法为查询用户 A 某天在该场所的所有按时间升序排列的记录信息, 相邻两条记录间隔小于阈值 a 的时间计算在停留时间内, 否则删除前一条记录, 继续往下进行.

单从停留时间识别用户身份, 拟合不了节假日期间用户停留过长问题. 场所人员的活跃度能从另一方面揭示身份信息, 考虑全天的活跃次数依然不好区分工作人员和顾客, 由于大型商业场所的营业时间与一般的场所情况不同, 研究商场的开业和歇业时间段出现的人群, 对区分工作人员和一般顾客有重要意义.

身份识别过程如下, 选择一个时间段, 计算历史用户在该时间段中停留的最大时间(阈值 b)和该时间段内开业和歇业时间区间出现的频次(阈值 c). 设定每天出现频次的最大值为 2, 即用户在开业和歇业时间区间都出现过记为 2, 只在一个区间出现过记为 1, 都未出现则记为 0, 对阈值 b 和阈值 c 的之间的关系进行回归分析, 提炼出回归公式, 其他用户根据回归公式进行身份识别.

1.2 小波神经网络模型

小波即为有限长度均值为 0 的波形, 小波变换是通过一个基本小波函数平移和伸缩构成一簇小波函数系去表达和逼近一个函数. 小波神经网络的隐含层传递函数为小波基函数, 信号前向传播而误差反向传播, 其三层拓扑结构如图 1 所示.

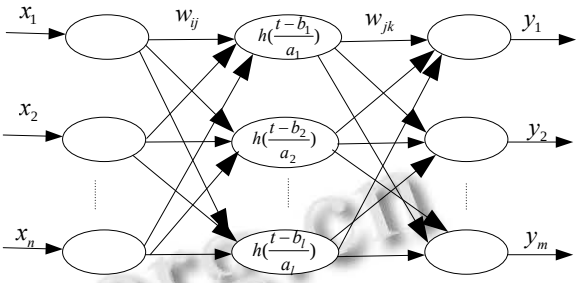

图 1　小波神经网络模型

其中输入为 $x_1,x_2,...,x_n$ ， $y_1,y_2,...,y_m$ 为预测输出, w_{ij},w_{jk} 为神经网络权值, 输入层、隐含层、输出层节点数分别为 n,l,m . 由于 Morlet 小波基函数在时、频域都有良好的局部特性, 且在支撑区域外快速衰减使得神经网络收敛速度加快, 它是对称的, 在二维情况下具有信号方向选择能力, 非常适合作为隐含层节点的传递函数, 其图形如图 2 所示, 表达式如下:

$$h(t)=\cos(1.75t)\exp(-t^2/2) \tag{1}$$

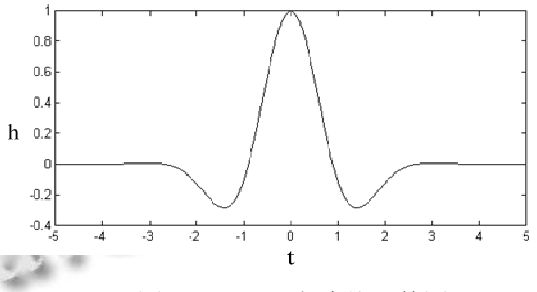

图 2　Morlet 小波基函数图形

小波神经网络学习训练的具体步骤如下:

步骤 1: 初始化网络参数 w_{ij},w_{jk},a_j,b_j , 其中 a_j,b_j 分别为小波函数的伸缩因子和平移因子, 网络参数通常在(0,1)区间上随机产生.

步骤 2: 输入训练样本并计算期望输出, 当输入为 $X_a=(x_{a1},x_{a2},...,x_{an})$ 时, 隐含层输出为:

$$h_a(j)=h_j\left(\frac{\sum_{i=1}^{n}w_{ij}x_{ai}-b_j}{a_j}\right) \qquad j=1,2,...,l \tag{2}$$

输出层计算公式为:

$$Y_a(k)=\sum_{j=1}^{l}w_{jk}h_a(j) \qquad k=1,2,...,m \tag{3}$$

步骤 3: 根据误差函数计算误差, 并采用梯度下降法修正网络权值和小波函数参数.

$$\text{误差函数:}\quad E=\frac{1}{2}(Y_a(k)-D_a(k))^2 \tag{4}$$

其中 $D_a(k)$ 为期望输出.

权值修正公式为:

$$w_{ij}(t+1)=w_{ij}(t)-\eta\frac{\partial E}{\partial w_{ij}}+\alpha\Delta w_{ij}(t+1) \tag{5}$$

$$w_{jk}(t+1)=w_{jk}(t)-\eta\frac{\partial E}{\partial w_{jk}}+\alpha\Delta w_{jk}(t+1) \tag{6}$$

$$a_j(t+1)=a_j(t)-\eta\frac{\partial E}{\partial a_j}+\alpha\Delta a_j(t+1) \tag{7}$$

$$b_j(t+1)=b_j(t)-\eta\frac{\partial E}{\partial b_j}+\alpha\Delta b_j(t+1) \tag{8}$$

式中 t 为当前训练次数, η 为学习速率, α 为动量因子.

步骤 4: 判断误差是否小于预先设定的某个值, 反之则返回步骤 2 , 输入下一样本. 当样本均循环一遍, 则进行下轮迭代, 直到误差满足条件或达到设定的迭代次数, 则停止网络的学习.

基于用户行为模型的客流量分析与预测总体流程图如图 3 所示:

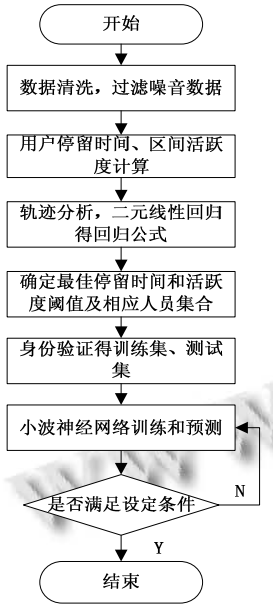

图 3 客流量预测总流程图

2 实验仿真与结果分析

2.1 数据预处理

数据来自南京某超大型商场的 2013 年 10 月 30 日至 2013 年 11 月 28 日共 30 天 9 个 AP 下的 Wi-Fi 数据. 本文只对 Mac 数据进行聚类分析, 不对单个 Mac 地址的隐私问题进行挖掘. 总数据量为 21,037,689 条, 其中独立的 Mac 地址记录 1,318,016 条, 考虑到数据量大和身份识别方案针对工作人员, 其岗位流动性较小等特点, 方案采取用 11 月 4 日开始的 7 天历史数据用于识别 Mac 用户身份. 根据停留时间算法, 获取次数小于 2 时, 停留时间为 0, 过滤掉此种噪音数据, 减少方案的计算量, 数据统计如表 1 所示, 并做出如下设定:

设定一: 基于 AP 日志生成和发送规则, 相邻两条日志记录间隔阈值 a 设为 5 分钟. 即 5 分钟内扫描到用户记录大于等于两次, 判断用户还停留在该场所, 间隔时间记录在停留时间内.

设定二: 根据商场的营业时间, 周日至周四为 10:00-22:00, 周五至周六为 10:00-22:30, 活跃度区间设为 09:00-10:30 和 21:30-23:00.

表 1 2013-11-04 至 2013-11-10 商场人数统计

	记录总条数	独立 Mac 数	过滤后 Mac 数
11.4	185,988	52,715	40,220
11.5	196,860	55,358	42,738
11.6	197,980	55,086	42,597
11.7	205,733	57,128	44,295
11.8	223,209	62,574	48,175
11.9	205,759	55,261	43,877
11.1	168,376	47,960	37,956

2.2 二元回归分析

根据上述设定, 计算用户在这七天中的停留时间和活跃次数. 由于工作人员数量较小, 方案的验证为根据实地调研确定人数区间范围, 其次查询用户的历史轨迹信息和在该 7 天在商场出现的总次数确定工作人员集合为 1365 人. 考虑到最大停留时间大于 4.5 小时的人数和活跃度次数大于 3 的分别占总数的 0.85%、1.1%, 则 b 为 min(*maxTime*,4.5) , c 为 min(*totleFrequency*,3) , 相互关系如图 4 所示.

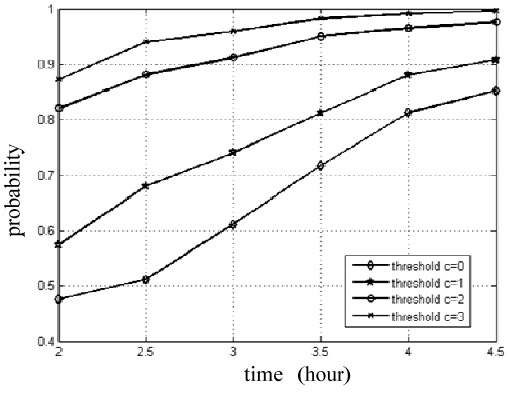

图 4 时间阈值与活跃度阈值相互关系

Matlab 多元回归分析函数 *regress*() 能够很好的求回归系数的点估计和区间估计，并检验回归模型，调用形式为：$[b,bint,r,rint,stats]=regress(Y,X,alpha)$，其中 *bint* 表示回归系数的区间估计，*r* 表示残差，*rint* 表示置信区间，*alpha* 表示显著性水平，*stats* 表示用于检验回归模型的统计量：相关系数 r^2、F 值、与 F 对应的概率 p，当相关系数 r^2 越接近 1、F 越大时，回归方程越显著；与 F 对应的概率 $p<alpha$ 时拒绝 H_0，回归模型成立. 由图 4 作置信水平为 95%的二元线性回归分析，可得如下公式:

$$y=0.0990b+0.0978c+0.3480 \quad (9)$$

其中相关系数 $r^2=0.86$，$F=71.659$，与 F 对应的概率 $p=1.3155e\text{-}10$ 小于显著性水平 0.05，回归模型成立，回归方程显著. 从图 5 所示的残差图可以看出，数据的残差离零点均较近，残差的置信区间均包含零点，说明回归模型能较好的符合原始数据.

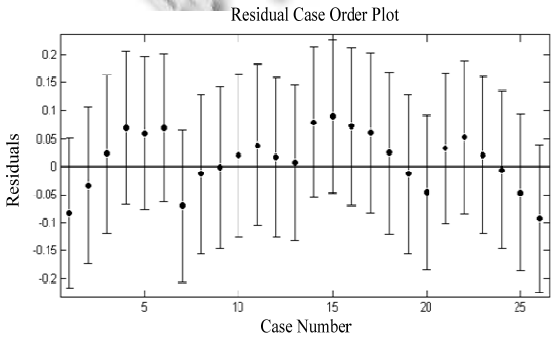

图 5　二元线性回归残差图

2.3 合理性验证

采用回归公式对 2013 年 11 月 11 日 14:00-15:00 一小时数量为 4236 的独立 Mac 用户身份进行抽样验证，结果如表 2 所示.

表 2　回归方程抽样验证分析

	识别人数	实际人数	相对误差
工作人员	301	275	9.45%
顾客	3935	3961	0.66%

结果显示在区分工作人员后，顾客的相对误差仅为 0.66%，相对误差较小，对本文的客流量预测非常有利. 根据轨迹分析，工作人员识别误差主要来自周边具有相同营业性质工作人员，其上下班经过此处，有时停留消费.

选择识别精确度在 90%左右，并且确定人数与实际人数相差较小，相应阈值可选 $c=2,b=3$，确定工作人员人数为 1454，图 6 为分离出工作人员和顾客后，每隔 5 分钟记录一次区间流量，11 月 4 日至 11 月 10 日各时间段被 Wi-Fi 感知的平均客流量分布. 为了体现工作人员人数变化，图 6 采用双纵坐标图，左纵轴刻度为总人数和顾客人数，右纵轴刻度为工作人员人数.

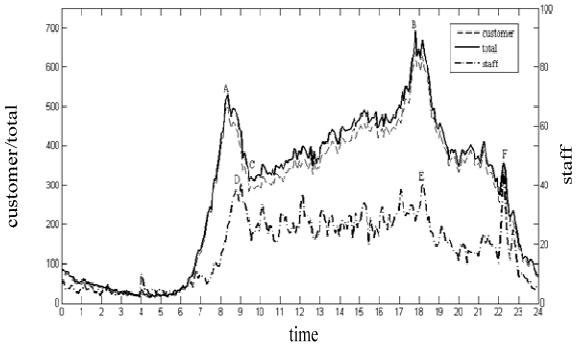

图 6　区间客流量分布

由图 6 可知顾客流量的 2 个峰值为 A(am 08:25,504)、B(pm 17:50, 662)，前后持续约 30 分钟，对应于上下班时间. 白天客流量从 C (am 09:35)持续上升到 B(pm 17:50)，而后急剧下降，晚上客流量主要集中于 pm 19:35- pm 21:40，而后逐渐减少；从工作人员波动曲线可知，白天在两个波峰 D(am 09:05)，E(pm 18:10)间波动，晚上的波峰出现在 F(pm 22:15)，与营业时间吻合. 凌晨的客流可能是 Wi-Fi 感应到的路人，而此时的员工人数可能是值班人员，其巡夜的过程会导致数据的变化. 人员流动模式与实际相符侧面反映回归公式的合理有效，同时表明选择区间 09:00-10:30 和 21:30-23:00 作为身份识别的正确性.

2.4 小波神经网络预测

根据回归公式，若要求身份识别精度在 90%左右，预测顾客的短时流量时，方案可以选择不同阈值确定的工作人员集合. 选择 2013 年 11 月 11 日至 14 日数据，采取每隔 5 分钟记录一次区间流量，数据为 4 天各个区间的平均值，并与根据历史轨迹信息确定工作人员集合情况进行比较，表 3 为不同阈值条件下对比情况.

表 3　不同阈值下误差对比情况

	确定人数	均方根误差
b=4.5,c=1	812	12.86
b=3, c=2	1454	7.72
b=2.5,c=3	1278	9.30

根据表 3 的比较结果可知: 选择$b=3,c=2$时, 均方根误差最小, 且确定人数与实际人数相当.

小波神经网络预测时, 采用 3 层神经网络, 输入层为当前时间的前 5 个时间点的客流量, 经网络训练后, 输出层输出当前时间的预测客流量, 隐藏节点数经多次比较选择 8 最合适, 采取每隔 5 分钟记录一次区间客流量. 训练集选择 2013 年 11 月 11 日至 13 日 3 天数据, 共 859 个样本, 测试集为 2013 年 11 月 14 日数据, 有 283 个样本. 学习速率为 0.01, 动量项学习速率为 0.01, 迭代次数为 800, 仿真平台为 Matlab R2012b, 训练结果采用 20 次实验结果的平均值. 图 7 为小波神经网络在$c=2,b=3$时客流量预测情况, 考虑图片清晰, 图 7 只展示了 am 10:00 至 pm 21:00 间预测情况, 由图 7 可知预测值和真实值非常接近, 为了评价预测准确度, 计算平均绝对误差百分比(MAPE)为 12.39%、均方根误差(RMSE)为 30.42, 对于这种长时间的短时流量预测, RMSE 较小, MAPE 小于 15%合理有效[10], 表明小波神经网络性能较好, 对此类短时流量问题能进行很好的预测.

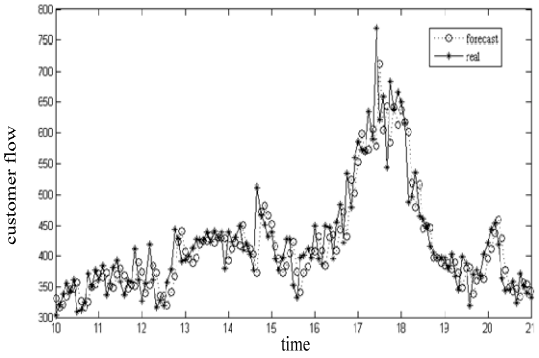

图 7 小波神经网络预测结果

3 结语

本文为了预测商圈的客流量, 提出将停留时间和特定时间段活跃度用于区分商场的工作人员和顾客的身份, 只对感兴趣的顾客流量进行预测, 并对拟合公式进行了合理性和有效性分析, 顾客的相对误差为 0.66%; 并讨论了不同阈值对预测的影响, 其中在给定$c=2,b=3$时, 利用小波神经网络进行短时客流量预测时性能较好, MAPE 小于 15%. 本文的工作对精准营销, 商业行为分析, 后勤保障等社会行为预测具有较广泛的应用意义. 下一步工作主要针对身份识别过程中对周边相同工作性质的人员区分以及小波神经网络这种梯度下降的局部搜索算法, 提出优化算法, 提高其预测精度.

参考文献

1 Lee S. Application of the Subset ARIMA Model for short-term freeway traffic volume forecasting. Transportation Research Record, 2009, 1678(35): 179–188.

2 王钰.基于小波神经网络的中国能源需求预测模型.系统科学与数学,2009,29(11):1542–1551.

3 Billings SA, Wei HL. A new class of wavelet network for nonlinear system identification. IEEE Trans. on Neural Networks, 2005, 16(4): 862–874.

4 冯再勇.小波神经网络与 BP 网络的比较研究及应用[学位论文].成都:成都理工大学,2007.

5 张坤,郁湧,李彤.小波神经网络在黄金价格预测中的应用.计算机工程与应用,2010,46(27):224–227.

6 任文君.基于网络用户行为分析的问题研究[学位论文].北京:北京邮电大学,2012.

7 何跃,陈大勇,腾格尔.基于 Web 数据挖掘的用户浏览兴趣路径研究.计算机工程与应用,2012,48(7):106–108.

8 王晓聪.基于位置的社交网络用户签到行为研究[学位论文].大连:大连海事大学,2012.

9 袁树寒,陈维斌,傅顺开.位置服务社交网络用户行为相似性研究.计算机应用,2012,32(2):322–325.

10 余国强.基于小波神经网络的短时交通流预测算法的研究[学位论文].广州:华南理工大学,2012.