

Study on ANN and ARIMA Combined Model for Short-time Passenger Volume Forecasting of Urban Rail Transit

Yan Zhou, Leishan Zhou, Yixiang Yue

School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China
Email: michelle5518@126.com, zhoulleishan@jtyu.edu.cn, yueyixiang@jtyu.edu.cn

Abstract: The result of the urban rail transit short-time passenger volume forecasting can promote urban rail transit operation management efficiency. According to the obvious real-time characteristics of passenger volume changes, the paper studies on the ARIMA and ANN models for predicting short-time passenger volume, and then raises an optimal combined model with ANN and ARIMA. Last, the application proves the validity and feasibility of the model and shows that the combined model has higher accuracy than the single prediction method.

Keywords: urban rail transit passenger volume forecasting; combined model; ARIMA; ANN

城市轨道交通短时客流 ANN 与 ARIMA 组合模型预测方法研究

周 妍, 周磊山, 乐逸祥

北京交通大学, 北京, 中国, 100044
Email: michelle5518@126.com, zhoulleishan@jtyu.edu.cn, yueyixiang@jtyu.edu.cn

【摘要】城市轨道交通短时客流预测结果将有效提高行车组织方案制定效率。本文根据城市轨道交通客流实时性变化特征, 分析 ANN 与 ARIMA 短时客流预测方法, 进而在两者基础上提出优化组合模型, 最后通过示例分析证明了模型的合理性, 且可有效提高短时客流预测的精确性。

【关键词】城市轨道交通短时客流预测; 组合模型; ARIMA; ANN

1 引言

城市轨道交通客流预测是制定轨道列车开行方案的基础。随着路网的形成, 已建设运营的线路客流分布将会随着网络中线路的增加发生变化, 网络路网中客流变化更具实时性特征。因此, 合理准确地预测网络化运营条件下轨道交通的客运需求, 将有效提高行车组织方案制定效率, 满足提高运营组织高效性和可靠性的需要。

目前我国城市轨道交通客流预测中, 普遍存在着规划阶段的预测结果与运营之后的实际客流有较大差异, 这主要因为网络轨道交通正处于积极建设阶段, 网络中各线路客流影响性和依托性逐步加大, 长期客流预测已难以准确反映客流的变化趋势。在轨道交通网络线路逐步建设运营时期, 短时客流预测将更实际可行的为行车组织方案编制服务, 提高运营组织效率。本文首先分析轨道交通客流特性, 阐述 ANN(神经网络)和 ARIMA(自

回归求和滑动平均)两种常用的短时客流预测方法^{[1][2]}, 并在此基础上提出 ANN 与 ARIMA 构建的组合预测模型, 最后通过示例分析, 证明了组合模型的可行性与精确性。

2 客流分析

2.1 客流在时间上的动态特性

1. 一日内小时客流分布具有不均衡性;
2. 一周内全日客流分布具有不均衡性;
3. 季节性或短期性客流的不均衡。

2.2 客流在空间上的动态特性

1. 线路具有方向动态性;
2. 断面客流具有动态性;
3. 各个车站乘降人数分布特征不一。

3 组合模型

3.1 ANN 模型

人工神经网络 (ANN) 建模方法是一种有效的分析预测方法。通常采用 BP 算法的多层前馈网络作为预测基础模型。1989 年 Robert Hecht-Nielson 证明了对于闭区间内的任一连续函数都可以用一个隐层的 BP 网络来逼近。以图 1 所示的三层前馈网络的应用最为普遍。

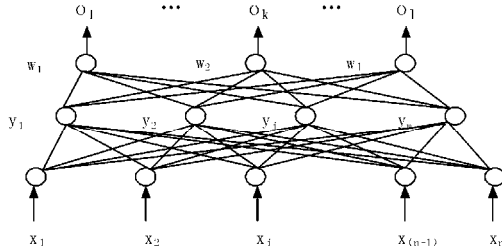


Figure 1. Three-tier feed forward BP neural network
图 1. 三层 BP 前馈网络

三层前馈网中, 输入向量 $X = (x_1, \dots, x_i, \dots, x_n)^T$; 隐层向量 $Y = (y_1, \dots, y_j, \dots, y_m)^T$; 输出层输出向量 $O = (o_1, \dots, o_k, \dots, o_l)^T$; 期望输出向量 $D = (d_1, \dots, d_k, \dots, d_l)^T$ 。输入层到隐层之间的权值 $V = (v_1, \dots, v_j, \dots, v_m)^T$, 其中 v_j 为隐层第 j 个神经元对应的权向量; 隐层到输出层之间的权值 $W = (w_1, \dots, w_k, \dots, w_l)^T$, 其中列向量 w_k 为输出层第 k 个神经元对应的权向量。

对于输出层, 有:

$$o_k = f(\text{net}_k), \text{net}_k = \sum_{j=1}^m w_{jk} y_j, k=1, 2, \dots, l \quad (2)$$

对于隐层, 有:

$$y_j = f(\text{net}_j), \text{net}_j = \sum_{i=1}^n v_{ij} x_i, j=1, 2, \dots, m \quad (3)$$

以上公式中, 转移函数 $f(x)$ 均为单极性 Sigmoid

函数 $f(x) = \frac{1}{1+e^{-x}}$, $f(x)$ 具有连续可导的特点, 且

$f(x) = f(x)[1-f(x)]$, 以上公式共同构成了三层前馈网的数学模型。式中各层权值调整公式为:

$$\Delta w_{jk} = \eta(d_k - o_k)o_k(1-o_k)y_j \quad (4)$$

$$\Delta v_{ij} = \eta \left(\sum_{k=1}^l (d_k - o_k)o_k(1-o_k)w_{jk} \right) y_j(1-y_j)x_i \quad (5)$$

常数 $\eta \in (0, 1)$ 为比例系数, 在训练中反映了学习速率。

3.2 ARIMA 模型

自回归求和滑动平均 (ARIMA) 模型是一种非平稳的时间序列模型, 它是参数化模型处理动态随机数

据的一种实用方法。其模型思想是对非平稳的时间序列用若干次差分使其成为平稳序列, 即差分次数为 d , 再用以 p, q 为参数的 ARIMA 模型对该平稳序列建模, 之后经反变换得到原序列。以 p, q, d 为参数的 ARIMA 模型预测方程可表示为:

$$y_t = \theta_0 + \varphi_1 y_{t-1} + \dots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q} \quad (6)$$

式中: y_t 为样本值; $\varphi_i (i=1, 2, \dots, p)$, $\theta_j (j=1, 2, \dots, q)$ 为模型参数; ε_t 为服从独立正态分布 $N(0, \sigma_{\text{noise}}^2)$ 的白噪声序列。 p, q, d 为模型的阶数, 通过对不同的 p, q, d 组合测试可以优化模型预测结果, 找到最合适的模型参数。

3.3 ARIMA 与 ANN 组合模型

为更有效的利用 ARIMA 模型与神经网络模型的优势, 依据 Bates 等在 1969 提出的组合预测思想, 即将参与组合的各种预测方法的结果通过适当的方式进行组合以获得最优预测结果的思想, 现建立 ANN 模型与 ARIMA 模型的组合模型, 以充分发挥两种模型的优势, 从而提高预测准确度。

模型思想: 引入非线性回归预测系数, 将通过已求得的非参数回归预测系数 K 计算出预测时间点客流量, 并将此预测结果与 ARIMA 模型此时间点客流预测结果共同作为神经网络模型的输入向量, 通过神经网络模拟, 输出此刻客流量的预测值。

这里非参数回归数学表达式:

$$\hat{y}_j^i = \frac{\sum_{i=1}^k y_j^i}{k} \quad (6)$$

其中 y_j^i 为预测时间 i 点之前天同时间 j 点的客流量。

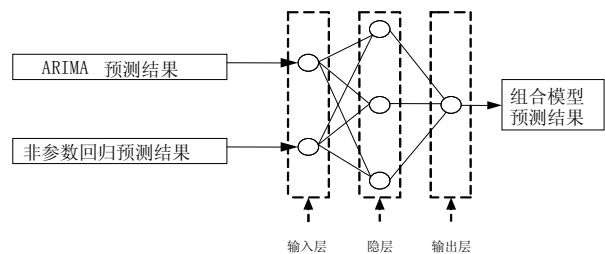


图 2. 组合模型结构

Figure 2. Combined model structure

则 ARIMA 与 ANN 组合模型预测共有 3 个步骤。

步骤 1: 计算预测时间点 t 的非参数回归系数 K , 预测出非参数回归预测结果 y_t^i 。

步骤 2: 用 ARIMA 模型对 y_t 进行预测。得到预测结果 y_t'' 。

步骤 3: 将 y_t' , y_t'' 输入神经网络输入层, 已知输入层为 2 个神经元, 输出层为 1 个神经元, 根据确定隐层节点个数选择经验公式来确定: $y_m = \sqrt{x(o+3)} + 1$, 其中 x , o 为输入输出节点数, y_m 为隐层节点数上限。这里取上限 3 个神经元, 以达到最佳预测结果。

这就是 ARIMA 与 ANN 组合模型的最最终预测结果。从预测过程看, ARIMA 模型将较优的预测出线性变化趋势下的客流预测结果, 非参数回归引入了近期客流变化趋势的预测, 两部分共同作用较好的降低了 ANN 模型训练难度, 在通过 ANN 模型的模拟预测将得到更为精准的预测结果。通过对 ARIMA 模型与 ANN 组合模型的综合运用, 发挥了这两种模型的长处, 克服了 ARIMA 模型对变化趋势变化波动较大的客流数据预测准确率低和 ANN 模型预测神经元个数选择难、计算复杂的问题, 提高了预测模型的可操作性和准确度。

4 示例计算

现以立水桥地铁站每小时入站客流量为实验数据, 预测下一时间段入站客流量为例, 演示各模型预测方法, 并分析实验数据。

模型 1 使用 ARIMA 模型进行预测, 根据模型思想应用 spss 软件中 ARIMA 程序计算。其运算过程为, 将 5 月 1 日至 18 日的客流数据, 以日为单位顺序排列成时间序列, 先对数据作平滑处理即将相邻序列数据作二次差分处理, 后根据公式 (1) 进行预测计算。

模型 2 使用神经网络模型, 同模型 1 先将客流数据排列成时间序列。后确定网络输入节点, 其方法是: 首先将序列分为两部分, 5 月 1 日至 15 日客流数据作为训练数据, 5 月 16 日至 18 日客流数据作为检验数据, 然后分别由小到大改变输入节点数来训练并检验其误差大小。当输入节点数增加, 误差不再有明显减小时, 则此时输入节点数变化的临界值即为应采用的数值。在根据经验公式计算出的隐层神经网络个数的范围内, 通过反复试算确定隐层个数, 为简化实验难度本实验中设定隐层节点个数为 4。

模型 3 不同于以上两个模型, 模型 3 先将对 5 月 1 日至 16 日同时间段上的客流依据公式 (6) 作循环测试, 找到各时间段上的最优非参数回归系数。用寻求到的非参数回归系数计算出 17, 18 日两天各时间段的客流量, 记录数值作为组合模型中神经网络的一个输入元数据, 另用模型 1 预测出同时间点客流量作为神经网络另一输入元数据, 设定隐层最大神经元个数 3, 用 17, 18 日实际客流量模拟调试组合模型神经网络, 最后重新计算 19, 20 日计算参数, 对 19, 20 日进站客流进行预测。

5 结果分析

5.1 模型评价标准

(1) 百分比绝对误差

表 1. 立水桥站小时进站客流量

Table 1. Hourly entrance in LiShuiQiao station

A站进站客流量 (人次)	05.01	05.02	05.03	05.04	05.05	05.06	05.07	05.08	05.09	05.10	05.11	05.12	05.13	05.14	05.15	05.16	05.17	05.18	05.19	05.20
5:00	143	122	143	107	131	142	137	163	142	135	120	112	138	124	120	140	124	130	116	128
6:00	684	643	664	633	654	674	643	678	633	654	674	643	678	633	654	674	643	678	633	654
7:00	3237	3185	3134	3119	3135	3206	3128	3150	3119	3135	3206	3128	3150	3119	3135	3206	3128	3150	3119	3135
8:00	2765	2728	2744	2703	2729	2756	2719	2746	2703	2729	2756	2719	2746	2703	2729	2756	2719	2746	2703	2729
9:00	1573	1475	1434	1465	1495	1523	1475	1468	1465	1495	1523	1475	1468	1465	1495	1523	1475	1468	1465	1495
10:00	1239	1223	1230	1210	1236	1276	1227	1260	1210	1236	1276	1227	1260	1210	1236	1276	1227	1260	1210	1236
11:00	1048	1075	1075	1034	1058	1068	1049	1077	1034	1058	1068	1049	1077	1034	1058	1068	1049	1077	1034	1058
12:00	1128	1135	1453	1118	1136	1150	1128	1163	1118	1136	1150	1128	1163	1118	1136	1150	1128	1163	1118	1136
13:00	1294	1243	1197	1237	1259	1288	1249	1287	1237	1259	1288	1249	1287	1237	1259	1288	1249	1287	1237	1259
14:00	1558	1456	1484	1486	1501	1534	1496	1534	1486	1501	1534	1496	1534	1486	1501	1534	1496	1534	1486	1501
15:00	1794	1756	1679	1704	1720	1752	1716	1745	1704	1720	1752	1716	1745	1704	1720	1752	1716	1745	1704	1720
16:00	2469	2336	2416	2417	2402	2431	2398	2434	2417	2402	2431	2398	2434	2417	2402	2431	2398	2434	2417	2402
17:00	3627	3586	3705	3564	3588	3604	3573	3706	3564	3588	3604	3573	3706	3564	3588	3604	3573	3706	3564	3588
18:00	2185	2175	2126	2150	2144	2154	2139	2173	2150	2144	2154	2139	2173	2150	2144	2154	2139	2173	2150	2144
19:00	1447	1446	1274	1430	1412	1463	1406	1456	1430	1412	1463	1406	1456	1430	1412	1463	1406	1456	1430	1412
20:00	1256	1275	1234	1243	1262	1286	1257	1265	1243	1262	1286	1257	1265	1243	1262	1286	1257	1265	1243	1262
21:00	1294	1149	1126	1156	1180	1234	1125	1172	1133	1179	1226	1145	1186	1135	1176	1206	1169	1196	1156	1180
22:00	638	636	653	647	639	680	633	653	624	641	692	655	638	604	622	680	646	653	650	649

$$E_{PAE}(t) = \frac{|y_t - \hat{y}_t|}{y_t} \times 100\% \quad (7)$$

(2) 平均百分比绝对误差

$$E_{MPAE} = \frac{1}{n} \sum_{t=1}^n E_{PAE}(t) \quad (8)$$

(3) 相关系数

$$r^2 = \frac{n \sum_{t=1}^n y_t \hat{y}_t - \sum_{t=1}^n y_t \sum_{t=1}^n \hat{y}_t}{\sqrt{(n \sum_{t=1}^n y_t^2 - (\sum_{t=1}^n y_t)^2)(n \sum_{t=1}^n \hat{y}_t^2 - (\sum_{t=1}^n \hat{y}_t)^2)}} \quad (9)$$

5.2 计算结果分析

由图 3, 4 和表 2 可见: 3 种预测方法中组合模型的 E_{MPAE} 最小, 相关系数 r^2 最大, 这表明组合模型的预测效果最好, 神经网络模型次之; 对非平稳、随机性较强的交通流预测时间序列, ARIMA 模型预测效果较差。

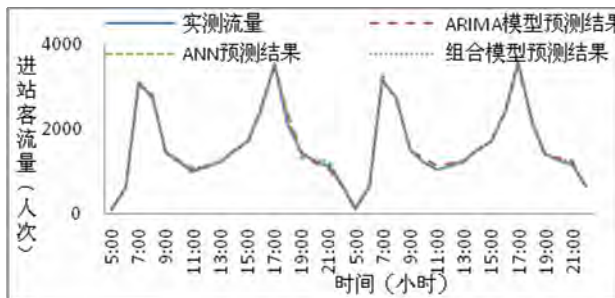
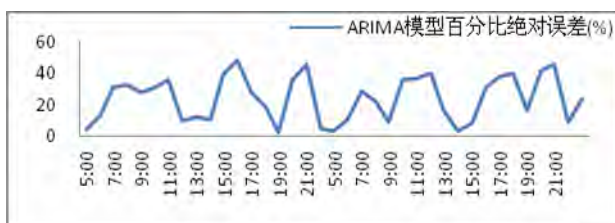
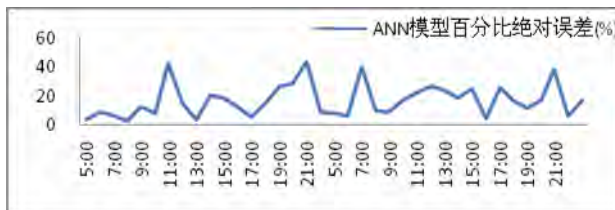


Figure 3. Compare between three predicted results with the actual data

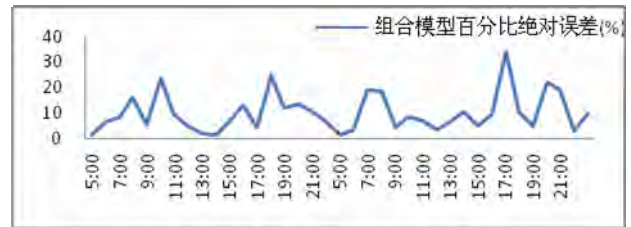
图 3. 三种预测结果与实时数据的比较



(a)The absolute proportion inaccuracy using ARIMA model
(a)ARIMA 模型的百分比绝对误差



(b)The absolute proportion inaccuracy using ANN model
(b)神经网络模型的百分比绝对误差



(c)The absolute proportion inaccuracy using combined model
(c)组合模型的百分比绝对误差

Figure4. The absolute proportion inaccuracy using the three kinds of models

图 4. 三种模型的百分比绝对误差

Table 2. Contrast of the average absolute proportion inaccuracy using three kinds of models

表 2. 三种方法的平均百分比绝对误差和相关系数对比

方法	ARIMA 模型	ANN 模型	组合模型
E_{MPAE}	26.483%	18.635%	11.393%
r^2	0.7242	0.7873	0.8064

6 结论

在短时轨道交通客流预测问题中, 由于交通流量的随机性和非线性程度较高, 单纯利用某一种预测方法难以得到满意的预测结果。本文应用 ARIMA 与 ANN 的组合模型对交通流量的实测数据进行预测与验证, 得到了较好的效果。这两种模型构建的组合模型为网络化轨道交通客流数据预测提供了新的方法。但由于组合模型比单一模型的计算量大, 当预测数据集进一步扩大时, 将加大预测问题的难度。

根据网络中客流的特点, 以预测客流量为依据, 制定轨道交通运营组织计划及相应的行车策略, 将是优化轨道交通运营管理、提高运营效益的有效手段。同时随着客流数据不断补充, 客流预测结果又将是判定轨道交通运营组织是否合理的评价标准。实时、可靠的客流预测结果不但可以反映轨道交通系统的运营现状, 还可对突发客流, 集聚客流等非正常客流现象出现趋势得以体现, 并对轨道交通运营管理起到警示的作用, 为运营组织对非正常客流的运输提供准备时间, 有效减少运营损失。

References (参考文献)

- [1] Jizhen Guan, Pengjian Shang. Non-linear Analysis and Forecasting of Traffic Flow [J]. Journal of System Science and Information, 2005(1):207-218.
- [2] Qinghuai Liang. An Analysis and Suggestions on Problems of Metro Traffic Volume Forecast [J]. Urban Rapid Rail Transits, 2005, 18(10), P23-30.
梁青槐, 城市轨道交通客流预测问题分析及建议[J], 城市轨道交通, 2005, 18(10), P23-30.