第三章 实验数据分析

3.1 数据分析

本文实验数据采集自武汉市友谊大道与园林路交叉路口 N31831G 地磁采集器。交叉路口设计图如图 3-1 所示。数据集包括从 2018 年 4 月 1 日至 2018 年 4 月 30 日交叉路口自西向东直行的所有车流量数据，每日数据采集从 0:00 至 24:00，间隔 15min，可得到 96 组数据，30 天共 2880($l = 2880$)条数据。按时间顺序对数据进行编号，纵坐标为车流量，原始数据散点图如图 3-2 所示。~~交叉路口设计图如图 3-1 所示。~~

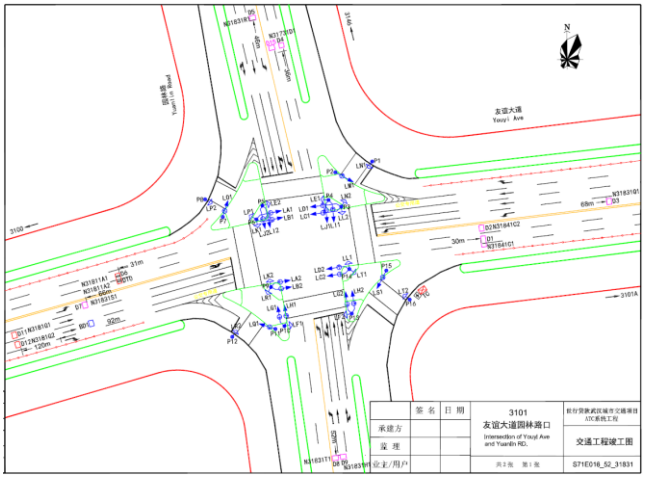

图 3-1 友谊大道与园林路交叉口设计图

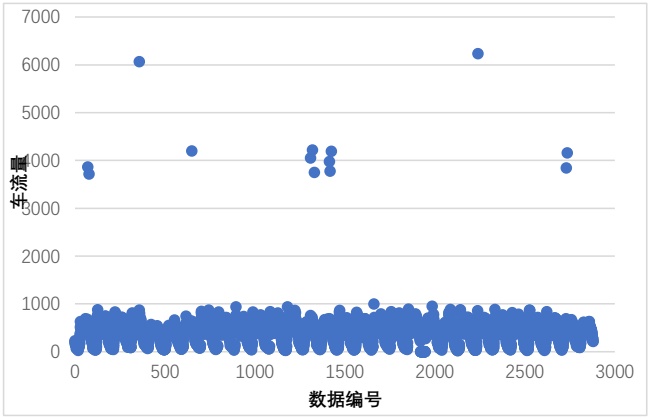

图 3-2 原始数据散点图（纵坐标单位）

3.2 数据修复

交通流数据在采集与储存过程中，由于天气或人为等干扰因素，容易出现采集设备、传输媒介或存储设备不稳定等问题，从而会导致收集到的数据有误。常见的数据问题可以分为两类：数据缺失与数据异常。从图 3-2 和对实验数据的观察可以看出，本文的实验数据也有数据缺失与数据异常问题出现。准确的观测数据是得到准确预测结果的前提，因此有必要对实验数据进行预处理。

对缺失数据，常见的补齐办法方法有历史趋势补齐、加权平均补齐和平均值补齐等。历史趋势补齐法是根据历史数据的发展趋势来估计缺失时刻的交通流；加权补齐法是对前一天的历史数据和缺失时刻的前一时刻的数据进行加权平均；平均值补齐法是对缺失时刻的前一时刻和后一时刻的交通流参数取平均[21]。

考虑实验数据的实际情况，论文分两种情况进行数据缺失的补齐考虑。

对单个的缺失数据，采用相邻时段的平均值补齐法修复数据，计算公式如下：

$$y(t) = round([y(t-1) + y(t+1)]/2) \quad (3\text{-}1)$$

连续多个数据缺失时，采加权补齐法，利用前一天同时段的历史趋势数据与需修复数据前一个时段的数据进行加权平均，计算公式：

$$y(t+1) = round\left(a * y(t-1) + (1-a) * y^{(k-1)}(t)\right) \quad (3\text{-}2)$$

其中，round()为四舍五入函数。a为加权系数，体现了历史趋势数据和当前数据在数据修复中所起的作用，数值越大表示当前数据对修复后的数据影响越大，论文取$a = 0.5$。$y(t-1)$为修复数据的前一个数据，$y^{(k-1)}(t)$为前一天与修复数据同时段的数据。

对存在异常的数据，即数据与不符合实际值的情况，采用阈值理论进行判断。阈值理论是根据对应的道路等级、车辆参数、控制类型等因素，设置交通流数据的合理区间，溢出区间的数值为错误数据[22][94]。车流量的合理区间计算公式为：

$$0 \le y \le f * C * T/60 \quad (3\text{-}3)$$

式中，y为车流量；f为修正系数，通常取 1.3-1.5；C为道路通行能力（veh/h），通常单车道通行能力取 2000veh/h（这里要有一个对应的速度等级或者道路等级的考虑，不同速度下通行能力是不一样的）；T为数据采集时间间隔。

由交叉路口设计图可知，直行车道为两车道，$T = 15\text{min}$，取$f = 1.3$，可得到：

$$0 \leq y \leq 1300 \quad (3\text{-}4)$$

故所收集的数据的合理区间为$[0,1300]$，超出此范围的数据视为异常数据，首先进行剔除，然后采用缺失数据的修复方法进行补齐。

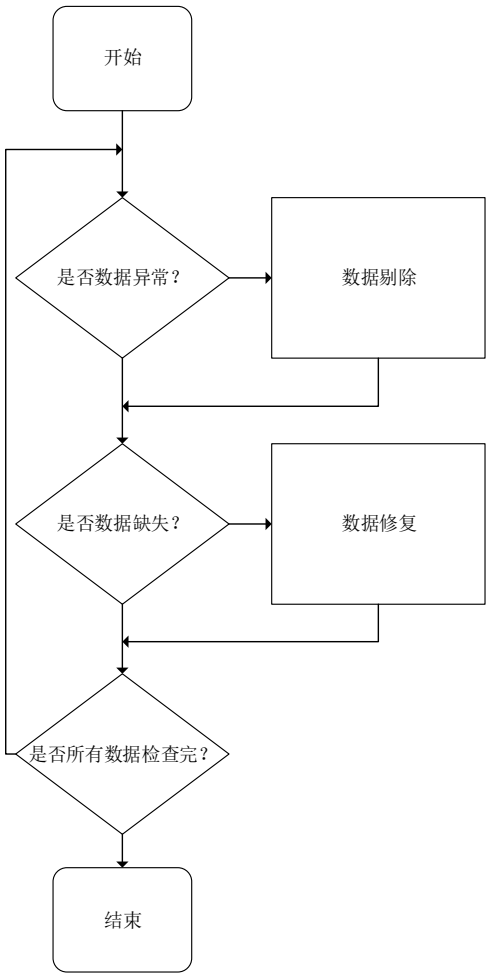

图 3-3 数据修复流程图

经过上述方法进行数据修复后，实验数据折线图如图 3-4 所示。

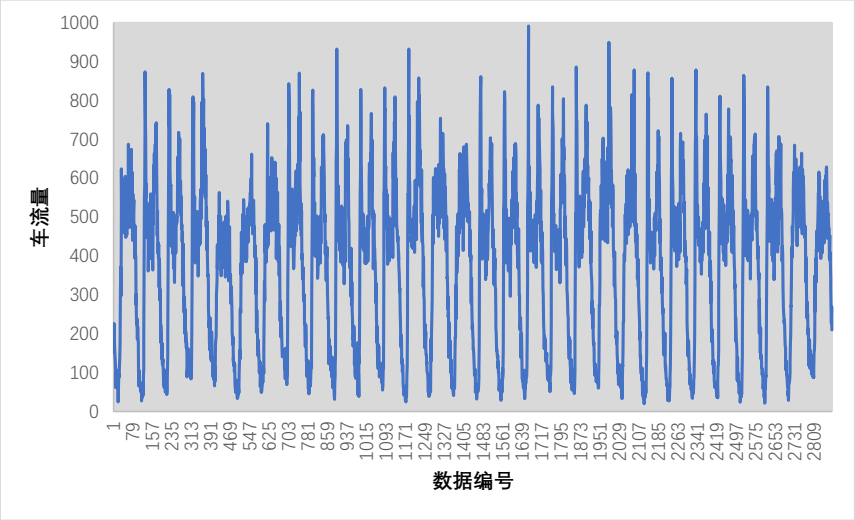

图 3-4 修复后数据折线图

四月第一周的车流量折线图如图 3-5 所示。可以看出，车流量在具有随机波动性的同时，也具有一定的规律性。每天车流量的变化趋势基本相同，车量高峰期集中在上午 7 点至 9 点之间和下午 5 点至 8 点间，这两个时间段之间的车流量也相对较高，车流量的低谷则分布在深夜 1 点至 5 点之间。

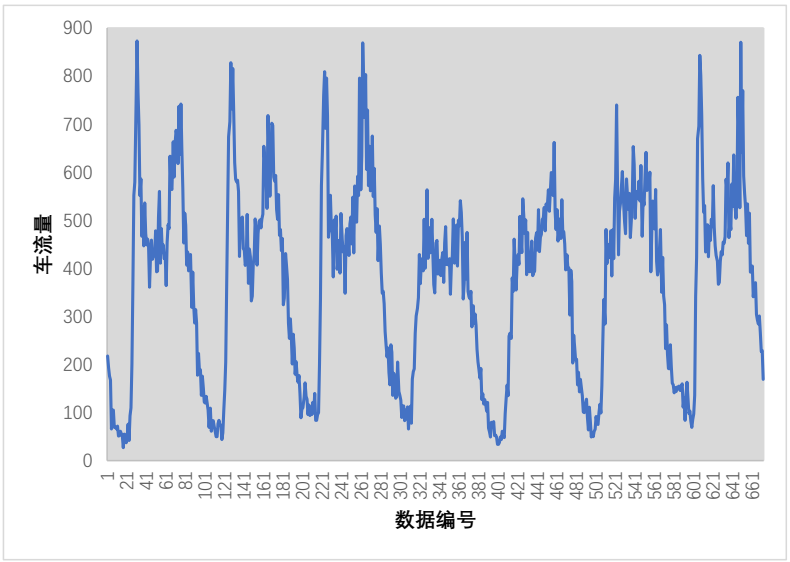

图 3-5 交叉路口车流量一周数据折线图

3.3 数据标准化

考虑到交通车流量数据在一天之内有较大的波动变化，上下班高峰期与深夜凌晨相差较大，数据整体分布分散。~~原始数据~~不同量级的原始数据容易对模型的鲁棒性造成影响，且支持向量机回归~~是~~建立在~~基于~~数据呈正态分布的假设之上~~数据，；~~再者，对数据~~标准化~~进行标准化处理可以提高算法的收敛速度~~，~~。因此，首先对数据集数据进行标准化处理，标准化公式如下：

$$z_i = \frac{x_i - \mu}{\delta} \tag{3-5}$$

其中x_i为i时刻的车流量，z_i为标准化后的数值，μ和δ分别为数据集的均值与方差。

3.4 评价指标

常见的预测性能评价指标有平均绝对百分比误差（Mean Absolute Percentage Error，MAPE）、平均绝对误差（Mean Absolute Error，MAE）、均方根误差（Root Mean Square Error，RMSE）、均等系数（Equal Coefficient，EC）。MAPE 表示误差占实际值的百分比，MAE 表示预测值与实际值的绝对误差，RMSE 表示预测值与实际值的均方根，EC 表示预测值与实际值的线性关系。前三个越小表示预测精度越高，EC 越接近于 1 表示预测值与实际值线性相关性越强，预测越精确。

$$\mathrm{MAPE} = \sum_{i=1}^{N} \frac{|\widehat{q_t} - q_t|}{q_t * N} \tag{3-6}$$

$$\mathrm{MAE} = \sum_{i=1}^{N} \frac{|\widehat{q_t} - q_t|}{N} \tag{3-7}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N} \frac{|\widehat{q_t} - q_t|}{N}} \tag{3-8}$$

$$\mathrm{EC} = 1 - \frac{\sqrt{\sum_{i=1}^{N} (\widehat{q_t} - q_t)^2}}{\sqrt{\sum_{i=1}^{N} \widehat{q_t}^2} + \sqrt{\sum_{i=1}^{N} q_t^2}} \tag{3-9}$$

其中$\widehat{q_t}$表示t时刻实际交通车流量值，q_t表示t时刻预测的交通车流量值，N为样本的数量。

第四章 基于 SVM 和 LSTM 的短时交通流预测

带格式的：突出显示

本章分别基于支持向量机和长短时记忆网络两种智能理论对短时交通流进行预测，并将预测结果与传统二次指数平滑法进行比较。

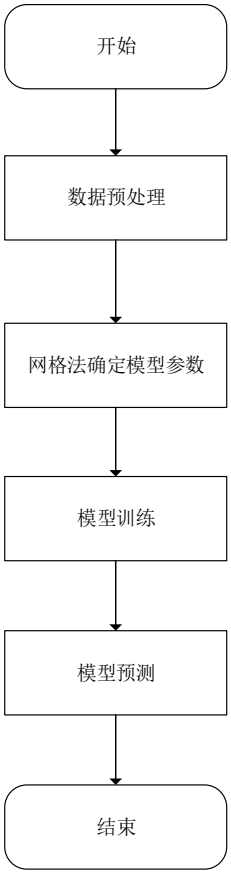

图 4-1 预测流程图

4.1 基于支持向量机的短时交通流预测

基于支持向量机回归模型，~~结合~~对武汉市友谊大道与园林路交叉口收集的车流量数据，进行短时交通车流量的预测。影响未来车流量的因素很多，天气、节假日、时间等，论文主要考虑前几个历史时刻的车流量对未来车流量的影响。

利用过去及当前车流量时间序列$\{y_1, y_2, \ldots, y_q\}$对未来第$s$个时刻的车流量$y_{q+s}$进行预测，其中$y_q$为第$q$个时刻的车流量；$q$为延迟阶数，即假设前$q$个时间段的车流量与未来预测有关；$s$为预测步长，当$s$=1 时，为单步预测，$s > 1$时，为多步预测。本文将收集到的 30 天数据，前 23（23*96=2208 条）天数据作为训练数据，后 7（672 条）天数据作为测试数据，用于检测预测模型的性能。

为了使数据适合支持向量机输入，需要重新构造数据集结构。给定数据集$D = \{(X_i, y_i)|i = 1,2, \ldots, l - q - s + 1\}$, 其中$X_i = (x_i, x_{i+1}, \ldots, x_{i+q-1})$为输入变量；$y_i = x_{i+q+s-1}$为预测输出值。得到的数据集结构如下：

$$X = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_{q-1} \\ x_2 & x_3 & x_4 & \cdots & x_q \\ x_3 & x_4 & x_5 & \cdots & x_{q+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{l-q-s+1} & x_{l-q-s+2} & x_{l-q-s+3} & \cdots & x_{l-s} \end{bmatrix} \quad (4\text{-}1)$$

$$Y = \begin{bmatrix} y_{q+s} \\ y_{q+s+1} \\ y_{q+s+2} \\ \vdots \\ y_l \end{bmatrix} \quad (4\text{-}2)$$

利用 python 编程语言，调用 sklearn 库中的 SVR 模块进行支持向量机回归预测，首先需要选择支持向量机的相关参数，包括核函数、正则化系数C、误差管道宽度ϵ等。合适的参数选择是进行准确预测的前提条件，对支持向量机参数的选择采用网格搜索法和经验法相结合的方法。常见的几种核函数在 python 中对应的参数如下表所示：

表 4-1 支持向量机参数表

核函数	正则化系数	误差管道宽度	多项式核函数系数	核系数	独立项
linear	C	epsilon			
poly	C	epsilon	degree	gamma	coef0
rbf	C	epsilon		gamma	
sigmoid	C	epsilon		gamma	coef0

基于~~对以~~往学者的经验，rbf 核函数是短时交通流预测使用最多且有效的核函数。因此本文支持向量机采用 rbf 核函数（高斯径向基函数），核函数的表达式为：

$$k(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) \quad (4\text{-}3)$$

对rbf核函数的三个参数，C、ϵ、γ采用网格搜索法进行确实，设置$C \in [0.1,10]$，$\epsilon \in [0.1,1]$，$\gamma \in [0.1,1]$，对三个参数的搜索步长都设为0.1，搜索过程以MAEP最小为目标。求解得到优化的参数$C = 1$，$\epsilon = 0.1$，$\gamma = 0.3$，网格法运行时间为15.21min。下图为利用网格搜索法求解三个参数的收敛图。

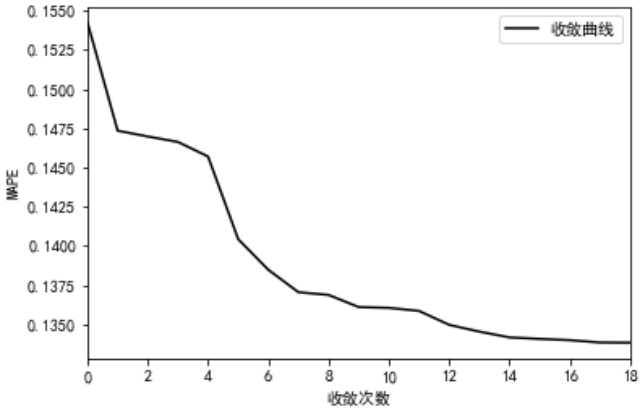

图 4-2 网格搜索法 MAPE 值收敛曲线

根据得到的参数，利用训练集中的数据训练支持向量机模型，并对测试集中的数据进行预测。取延迟阶数$q = 10$，预测步数$s = 1$，即依据历史前10个时间段的交通车流量，预测下一个时段的车流量。测试集预测得到的预测值与实际值如图4-3所示。

表 4-2 支持向量机参数设置表

参数	核函数	C	ϵ	γ	q	s
值	rbf	1	0.1	0.3	10	1

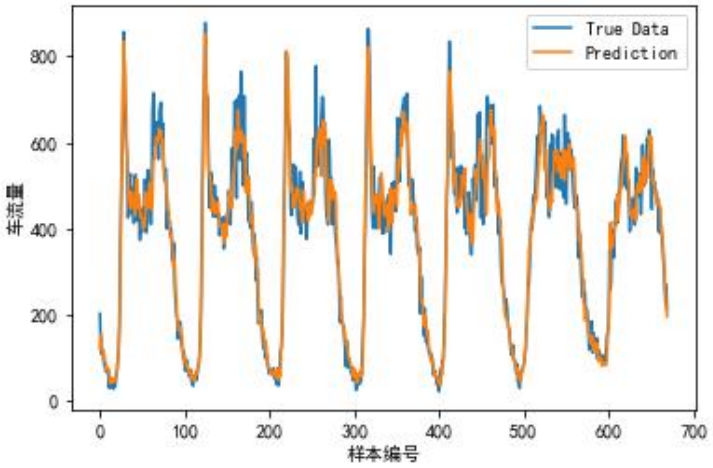

图 4-3 支持向量机预测对比图

图 4-4 是 4 月 24 日一天的预测值与实际真实值的折线图及真实值与预测值残差的散点图，可以更直观地看出预测效果。

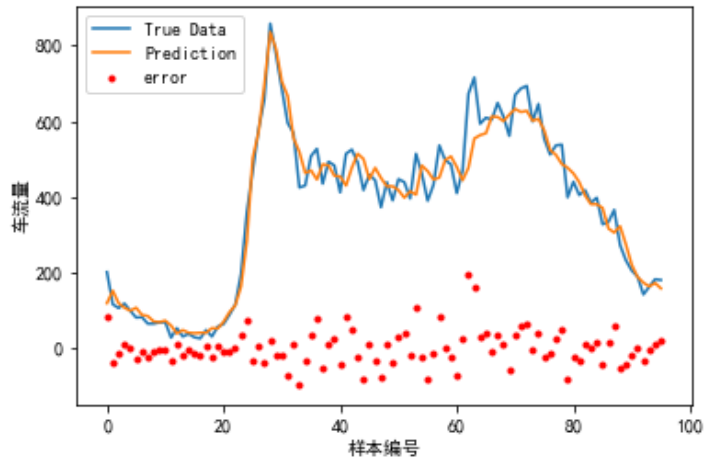

图 4-4 支持向量机一日预测对比图

从以上两张图可以看出，对测试集~~的~~进行拟合的曲线基本符合实际测试值数据的曲线~~，~~。在交通车流量变化较大时，依然保持较一致，从直观的角度看，预测较符合实际。由于交通车流量数据具有随机波动性，因此预测结果与实际值存

在一定残差，残差也表现出一定程度的随机波动性。

表 4-3 支持向量机预测评价指标值

评价指标	指标值
MAPE	13.39%
MAE	39.98
RMSE	54.76
EC	0.94

用 MAPE、MAE、RMSE、EC 四个指标来评价支持向量机回归预测的性能。得到四个指标如表 4-3,MAPE=13.39%，MAE=39.98，RMSE=54.76，EC=93.66%。由数据可以看出，预测的精度在 85%以上，并且预测值与实际值的线性关系也在 90%以上，说明支持向量机在短时交通流预测领域具有可行性，基于结构风险最小化的原则使得支持向量机回归模型具有较好的泛化能力，在遇到未知数据时仍能表示良好的预测性能。

4.2 基于长短时记忆网络的短时交通流预测

构建长短时记忆网络预测模型，使用与支持向量机回归预测相同的数据集、同样的标准化方法和评价指标，对短时交通流进行预测。

编写 python 程序，调用 keras 库中的 LSTM 模块构造长短时记忆网络预测模型。模型参数的选择对于提高预测精度具有重要的作用，LSTM 的参数可分为两类，一是模型在训练学习过程中自动调整的参数，如在第二章介绍中提到的权重和偏置；另一类是需要人工设定的参数，包括输入层、隐藏层、输出层的数量，各层神经节点的数量，激活函数等。

目前对人工参数的选择没有既定的选择方法，多采用试凑法或经验法进行调整。与支持向量机参数的选择方法类似，对长短时记忆网络采用网格搜索法与经验法结合的方法确定人工参数。考虑到时间成本和设备因素，对隐藏层的数量、隐含层节点的数量和暂停率三个重要的参数，编程通过网格法进行确定；对优化器、激活函数、损失函数等参数，采用经验法确定。

表 4-4 长短时记忆网络参数设置表

参数名称	参数取值
隐藏层数	2
隐藏层节点数（units）	35
暂停率（dropout）	0.2
输入步长（input_length）	10
批处理数量（batch_size）	100
训练期数（epochs）	10
优化器(optimizer)	rmsprop
损失函数（mse）	mse

表中为使用 keras 构建 LSTM 模型时，几个重要的人工调整的参数，括号内为参数在 python 语言中的符号表达。考虑到网格法程序运行时间，将每个隐藏层的节点数量设为相同。设置隐藏层数的范围在[1,5],步长为 1；隐藏层节点数设为[30,50]，步长为 5；暂停率是为了防止训练过程中过拟合，每次训练随机选择部分神经节点使其不工作，暂停率即是不工作神经节点占总节点数量的比率。设置搜索范围为[0.1,0.3]，搜索步长 0.1。

以减小 MAPE 值为目标，通过网格法搜索得到优化的参数为隐藏层数为 1，每层神经节点数为 40，暂停率为 0.1，搜索时间 41.84min，下图为网格搜索过程中 MAPE 值的变化。

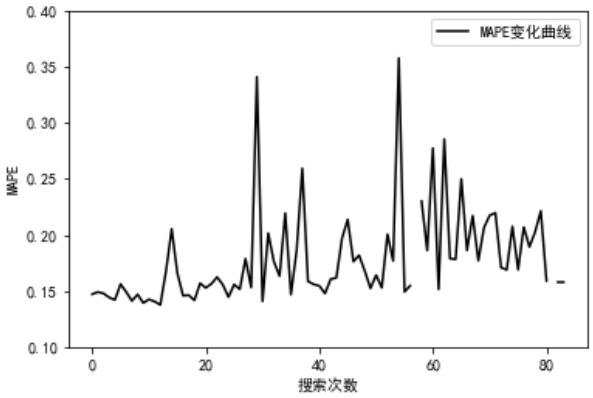

图 4-5 网格搜索法 MAPE 值变化曲线

LSTM 模型输入步长指自变量X的维度，即延迟阶数，这里设置与支持向量机回归相同的延迟阶数$q=10$，相同的预测步长$s=1$。

批处理数量指每次训练的样本数量，训练期数指训练完全部样本的次数。批处理数量较大可更好地把握总体样本的特征，但批处理数量大对 PC 内存要求较高，较小的批处理数量可通过增大训练期数来达到相同的预测效果，取批处理数量为 100，训练期数也为 10。

优化器是 LSTM 训练过程中的计算算法，损失函数相当于是优化器的目标函数。其中 rmsprop 优化器和 mse(均方差)损失函数是 LSTM 预测中使用较多的选择，本文也采用这种设置。

根据以上的参数设置，得到 LSTM 模型的总体结构如图所示。

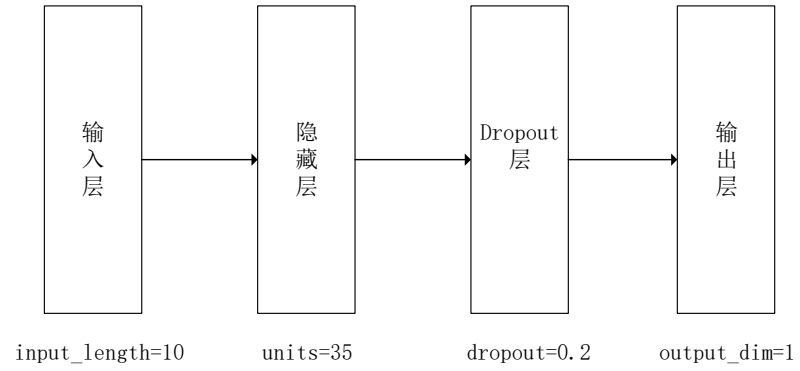

图 4-6 长短时记忆网络模型结构图

使用训练集里的数据训练构建的模型，用测试集的数据对模型性能进行评估。七天预测的车流量与实际车流量的折线图如图 4-7 所示，为了更直观观察预测的效果，将 24 日一天的预测结果和残差表示出来。

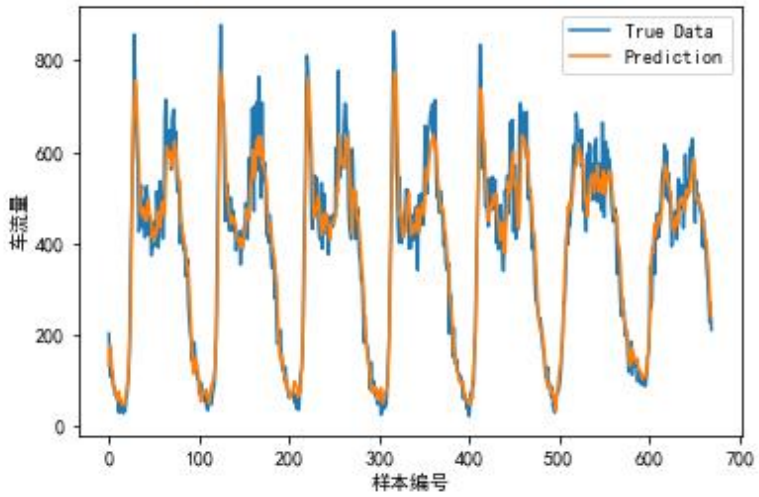

图 4-7 长短时记忆网络预测对比图

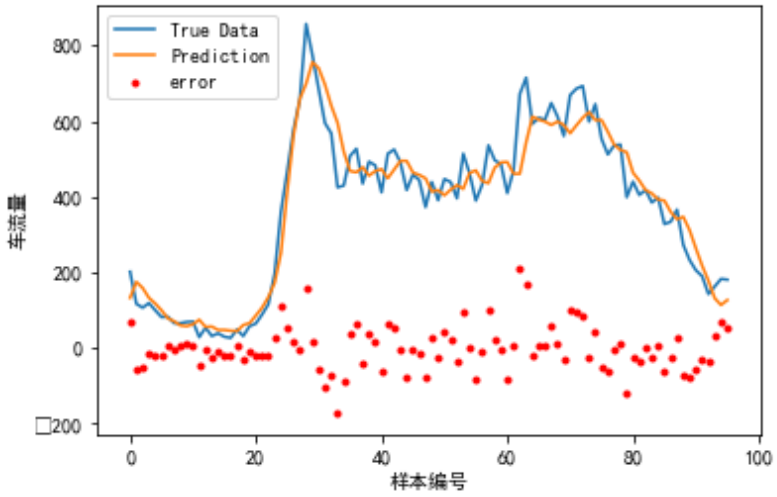

图 4-8 长短时记忆网络一日预测对比图

从两张图中可以看出，预测值在趋势上和数值上都与实际数据比较符合，大部分数据的预测误差都较小。

表 4-5 长短时记忆网络预测评价指标值

评价指标	指标值
MAPE	13.77%
MAE	44.12
RMSE	60.03
EC	0.93

上表为预测评价指标的值，MAPE=13.77%，MAE=44.12，RMSE=60.03，EC=0.93。从评价指标可以看出，利用 LSTM 模型对短时交通流进行预测的精度在 85%以上，绝对误差值较小，预测数据与实际数据的线性系数在 0.9 以上，预测性能良好。

说明了长短时记忆网络这种循环神经网络的变种，在处理较长时间的历史数据时，确实有较好的记忆能力，且不会出现梯度爆炸问题。再者，长短时记忆网络在短时交通流预测领域具有很好的可移植性和可行性。

4.3 与传统模型比较

为比较智能理论与传统理论在预测性能上的区别，将两种智能理论预测的结果，与传统的二次指数平滑法预测的结果进行比较。

设置平滑系数$\alpha = 0.5$，$q = 30$，$s = 1$，即每 30 个历史时段数据采用一次二次指数平滑法预测下一个时段的数据。以数据集的后 7 天作为预测对象。

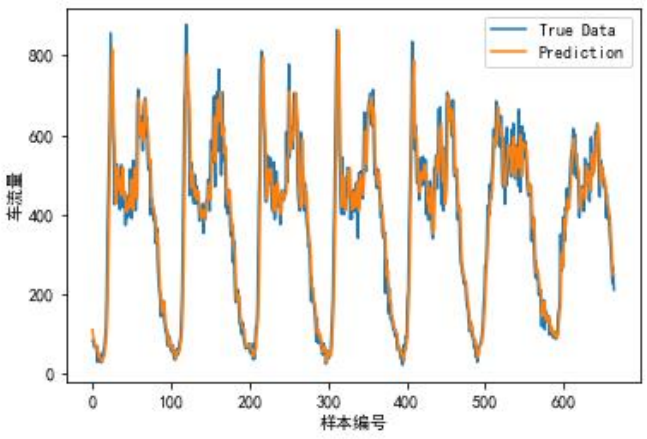

图 4-9 二次指数平滑法预测对比图

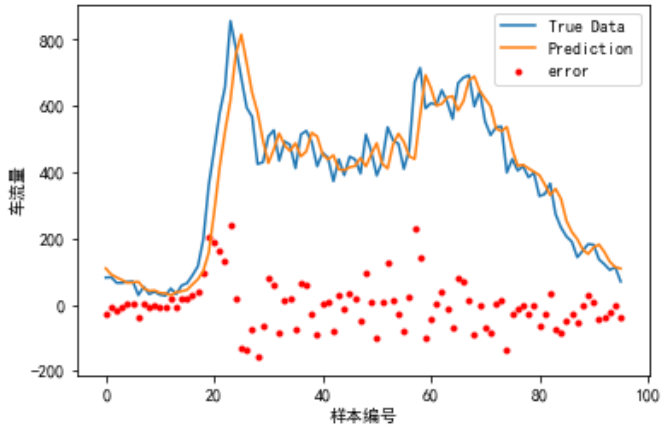

图 4-10 二次指数平滑法一日预测对比图

从总体图中可以看出，二次指数平滑预测的结果在总体变化趋势上基本符合车流量的变化趋势，但是从单日的图中可观察到，预测值相较于真实值有超前的现象。在车流量变化较大的时段，残差也较大。

表 4-6 预测指标比较

预测方法	MAPE	MAE	RMSE	EC
SVM	13.39%	39.98	54.76	0.94
LSTM	13.77%	44.12	60.03	0.93
二次指数平滑	16.39%	51.28	71.77	0.92

从表中可以看出，与传统二次指数平滑预测方法相比，SVM 和 LSTM 在四个指标上都明显更优。该现象说明，在短时交通流预测问题上，SVM 和 LSTM 的预测性能优于二次指数平滑法。智能理论预测方法相比于传统基于统计理论的预测方法，能够更好地抓住短时交通流变化的随机因素，并且在交通流变化较大时，仍能有较好地预测。

4.4 本章小结

本章首先对实验数据进行分析，采用阈值理论判断异常值，利用相邻时段补齐法和加权平均法进行缺失数据的修复，通过预处理将数据标准化，并设定预测

的评价指标。而后分别利用支持向量机和长短时记忆网络对短时交通车流量进行预测，采用网格搜索法确定两个模型的参数，得到符合预期的预测结果。最后将两种智能理论模型的预测结果与传统基于统计理论的二次指数平滑预测结果进行对比，得到智能理论模型在预测性能上要优于传统模型的结论。