学号 142006010324

年级 2014级



**本科毕业论文**

**基于多源数据的高速公路短时交通流量预测**

|  |  |
| --- | --- |
| **专 业** 计算机科学与技术 |  |
| **姓 名** 邵 程 立 |  |
| **指导教师** 谢 在 鹏 |  |
| **评 阅 人** |  |

**2018年5月**

**中国 南京**

**BACHELOR'S DEGREE THESIS**

**OF HOHAI UNIVERSITY**

**Short-term Traffic Flow Forecast Based on Multi-source Data**

College : College of Computer and Information

Subject : Computer Science and Technology

Name : ChengLi Shao

Directed by : ZaiPeng Xie Lecturer

NANJING CHINA

**郑 重 声 明**

本人呈交的毕业论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本设计（论文）的研究成果不包含他人享有著作权的内容。对本设计（论文）所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本设计（论文）的知识产权归属于培养单位。

本人签名： 日期：

摘 要

交通流量预测是一种间接预测交通状况的方法，实时的交通状态对交通控制和智能城市的研究有重大意义。交通流量预测依据时间长短划分为：短期预测、中期预测、长期预测。由于交通流量具有波动性，为了提高预测的精确度，研究短时交通流量更有意义。

短时交通流量预测可以看着是一种非线性回归问题，从历史数据中挖掘出有迹可循的规律或者说成是学习拟合函数的各种参数，并将其应用于未来数据的预测上。大量有效的数据是实现这一目标的基础。传统的有效数据一般仅为历史的交通流量，而忽视交通流量在在时间和空间上的相关性。为了提高交通流量预测的准确性，使用多源数据进行短时交通流量预测，增加各种相关因素，从而提高准确率。多源数据即多种相关联的数据，例如，测量点周围的环境因素（降雨量，温度，风速）。

已经有很多机器学习的算法应用于流量预测，本文采取当前比较流行的算法，如传统支持向量回归（SVR）算法，人工神经网络中的多层前馈神经网络算法，以及对时间序列有显著效果的循环神经网络LSTM算法，同时使用一定权值，将各个模型进行集成应用在针对高速公路的流量进行分析。

本文使用多源数据进行高速公路短期流量预测，实现了基于BP神经网络的多种回归算法相结合的模型构成的短期流量预测系统。短期交通流量预测的研究对智能交通系统的发展有重大意义，而预测的精确度更是衡量智能交通系统的指标之一。

**关键词** 交通状态；短时交通流量；多源数据；SVR；神经网络；集成学习

**ABSTRACT**

Traffic flow prediction is an indirect method to predict traffic condition, and the real-time traffic condition is of great significance for the study of traffic control and intelligent city. The forecast of traffic flow is divided into three types: short, medium and long term. Because of the volatility of traffic flow, it makes sense to study short-term traffic flow in order to improve the accuracy of prediction.

Short-term traffic flow prediction can be considered as a kind of nonlinear regression problems, excavated from the historical data has a track record of law or to learn the various parameters of the fitting function and applied to predict future data. A large number of valid data is the basis for achieving this goal. The traditional effective data is generally only the traffic flow history, but the correlation between time and space is ignored. In order to improve the accuracy of traffic flow prediction, multi-source data is used for short-term traffic flow prediction, and various related factors are added to improve the accuracy. Multi-source data is a variety of correlated data, such as the environmental data (rainfall, temperature, wind speed).

There are a lot of machine learning applications intraffic prediction, this paper compares the traditional popular support vector regression (SVR) with the artificial neural network including GA\_BP and LSTM algorithms. Certain weight will be simultaneously combined with each model. In order to reduce the complexity of the study, the traffic flow of the highway is analyzed.

The integration method is used to predict the traffic flow on the expressway by combining two high quality algorithms. The former combines the advantages of various methods, the latter simplifies the network structure and reduces the correlation factor.

This paper predict the short-term traffic flow of expressway, and realized the short-term flow prediction system based on the combination of multiple regression algorithms with BP neural network. The study of short-term traffic flow forecast is of great significance to develope the intelligent transportation system, and the accuracy of prediction is one of the focuses of intelligent transportation system.

**Keywords:** Traffic Status; Short-term Traffic Flow; Multi-source Data; SVR ; NN integrated learning ;

**目录**

[摘 要 I](#_Toc516435033)

[ABSTRACT II](#_Toc516435034)

[目录 III](#_Toc516435035)

[第1章 绪论 5](#_Toc516435036)

[**1.1研究背景** 5](#_Toc516435037)

[**1.2国内外研究现状** 7](#_Toc516435038)

[**1.3 本文章节介绍** 9](#_Toc516435039)

[第2章 TensorFlow框架介绍 11](#_Toc516435040)

[**2.1 TensorFlow简介** 11](#_Toc516435041)

[**2.2 交通状态** 13](#_Toc516435042)

[**2.3 小结** 13](#_Toc516435043)

[第3章 短时流量预测的相关算法 15](#_Toc516435044)

[**3.1 支持向量回归** 15](#_Toc516435045)

[**3.2 遗传算法** 18](#_Toc516435046)

[**3.3 BP神经网络** 20](#_Toc516435047)

[**3.4 长短期记忆网络** 23](#_Toc516435048)

[**3.5 小结** 28](#_Toc516435049)

[第4章 基于多源数据的高速公路流量预测系统设计 29](#_Toc516435050)

[**4.1 需求分析** 29](#_Toc516435051)

[**4.2 总体设计** 29](#_Toc516435052)

[**4.2.1 系统结构设计** 29](#_Toc516435053)

[**4.2.2 系统流程设计** 29](#_Toc516435054)

[**4.3详细设计** 32](#_Toc516435055)

[**4.3.1 处理输入数据** 32](#_Toc516435056)

[**4.3.2 定义模型参数** 32](#_Toc516435057)

[**4.3.3 确定模型评估方式** 33](#_Toc516435058)

[**4.3.4 界面设计** 34](#_Toc516435059)

[**4.4 小结** 34](#_Toc516435060)

[第5章 基于多源数据的高速公路流量预测系统实现 35](#_Toc516435061)

[**5.1 系统开发环境** 35](#_Toc516435062)

[**5.2样本提取** 35](#_Toc516435063)

[**5.2.1 数据准备** 35](#_Toc516435064)

[**5.2.2 工具介绍** 35](#_Toc516435065)

[**5.2.3 提取步骤** 35](#_Toc516435066)

[**5.2.4 提取步骤** 37](#_Toc516435067)

[**5.3 系统实现** 39](#_Toc516435068)

[**5.3.1 输入数据功能实现** 39](#_Toc516435069)

[**5.3.2 模型训练功能实现** 39](#_Toc516435070)

[**5.3.3 模型验证功能实现** 42](#_Toc516435071)

[**5.3.4 模型预测功能实现** 43](#_Toc516435072)

[**5.3.5 系统界面实现** 43](#_Toc516435073)

[**5.4 系统展示** 43](#_Toc516435074)

[**5.4.1 系统界面展示** 43](#_Toc516435075)

[**5.4.2 选择文件** 43](#_Toc516435076)

[**5.4.3 模型训练** 44](#_Toc516435077)

[**5.4.4 模型验证** 44](#_Toc516435078)

[**5.4.5 进行预测** 45](#_Toc516435079)

[**5.5 实验结果-数据** 45](#_Toc516435080)

[**5.6 实验结果-模型** 46](#_Toc516435081)

[**5.7 实验总结** 47](#_Toc516435082)

[**5.8 小结** 47](#_Toc516435083)

[第6章 总结与展望 49](#_Toc516435084)

[致 谢 51](#_Toc516435085)

[参考文献 52](#_Toc516435086)

**第1章 绪论**

**1.1研究背景**

近年来，随着生活质量的改善，人们对于外出的看法有了不一样的认识，开始追求精神上的享受，自驾旅游慢慢成了人们生活中不可或缺的部分，另一方面，随着汽车等交通工具的普及（据公安部统计，截至2017年底，全国机动车保有量达3亿辆，机动车驾驶人接近4亿人），驾车外出的情况也逐年爬升。相关数据显示，下面是机动车新用户趋势图[1]，这种情况还在不断加剧。

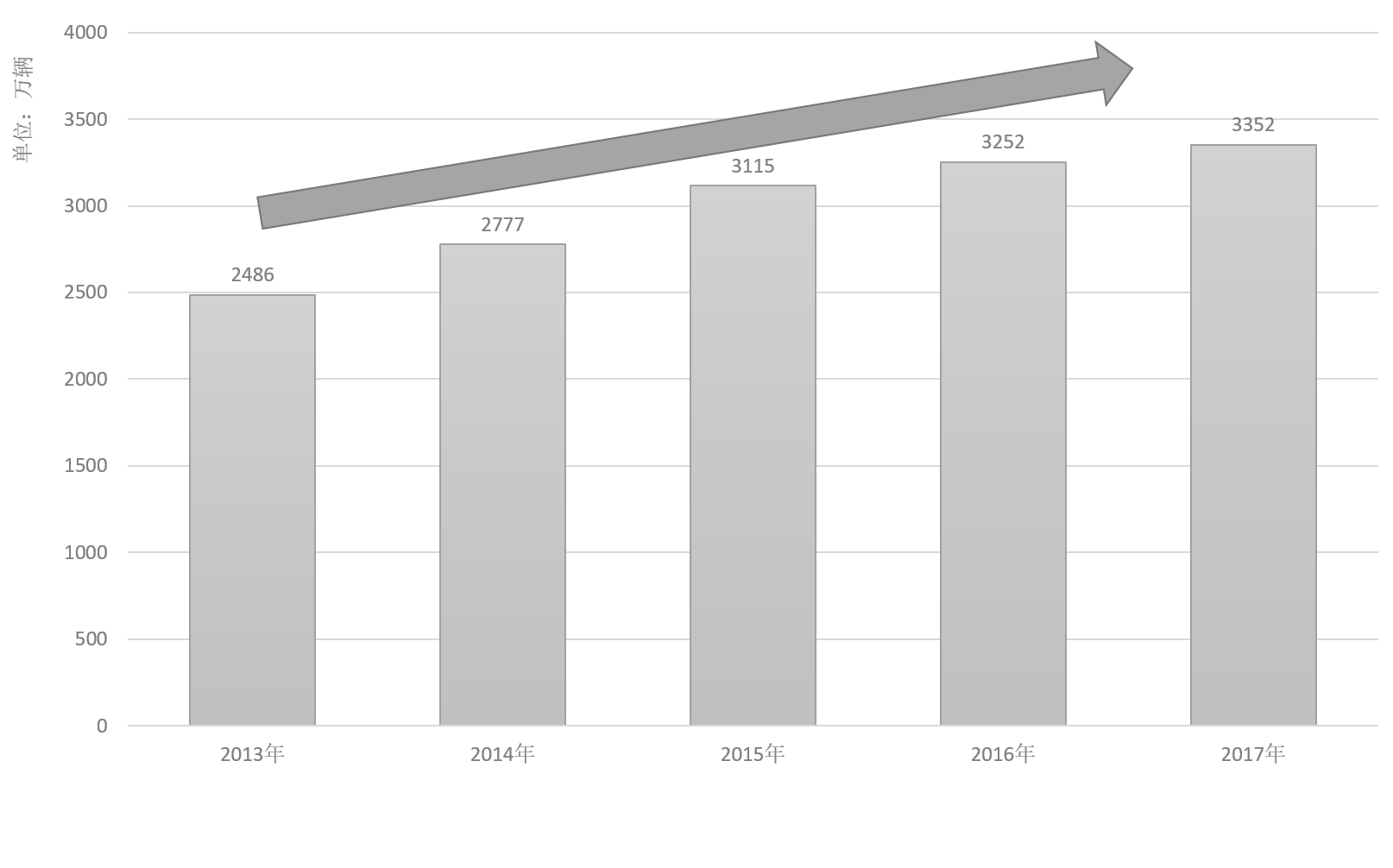


图1-1 机动车新用户趋势图

随着车辆数目的攀升，公路的负载也在不断增大，一些地方甚至已经出现超负载的情况。虽然我国公路也在不断发展（依据中国国家统计局的年度数据现实2016年公路里程469.63万公里，高速等级路公路里程13.10万公里）。但是严重的堵车现象还是在局部地区出现，看似和总的公路里程没有关联，但随着路网的不断增加，爆发点周围的路网在一定程度上应该能缓解这个病情，至于缓解的多少，主要看处理的方式，然而一般的处理方式，仅仅是增加人手进行维护秩序，以及严重堵车后进行限流处理，而在限制车流量和释放车流量的处理上存在严重不足，无法做到实时更新。与此同时，随着科技的发展，人们享受着愈加便利的生活方式，以及身处在较快的生活节奏之中，简洁、快速的生活理念扎根在人们心中，时间的重要性也越来越受到人们的关注。然而有时候，在一个十字路口，看到不同方向上的车或者人流量比自己这个方向的少的多，即十字路口交叉道路的效率严重不对等。造成人们对现有交通系统的不满。我国目前的交通系统大部分还停留在基于规则的系统之下，更确切的说是一个静态的交通系统。由于静态基于固定时间变化的交通指示灯无法根据实时的道路状况进行最优调度，致使公里资源没有充分被利用。毫无意义的等待致使人们开始抱怨，人们开始追求能动态调整的交通系统。

智能交通系统（ITS）的概念在人们的期待中上线，根据2018年的一篇来自腾讯互联网+的文章[2]，智能交通系统的概念如下：先进的科学技术，类似于万物互联的感觉，通过采集各种数据，提取有用的数据并进行处理得到某种结论，再在各个系统中将有效信息进行交换，最终每个部分都知道全局的信息，而不是局部的信息，利用掌握的全局信息和各种有效算法进行最优的调度，到达各种资源最优化利用。下面是城市智能交通系统管理平台的概率图[2]

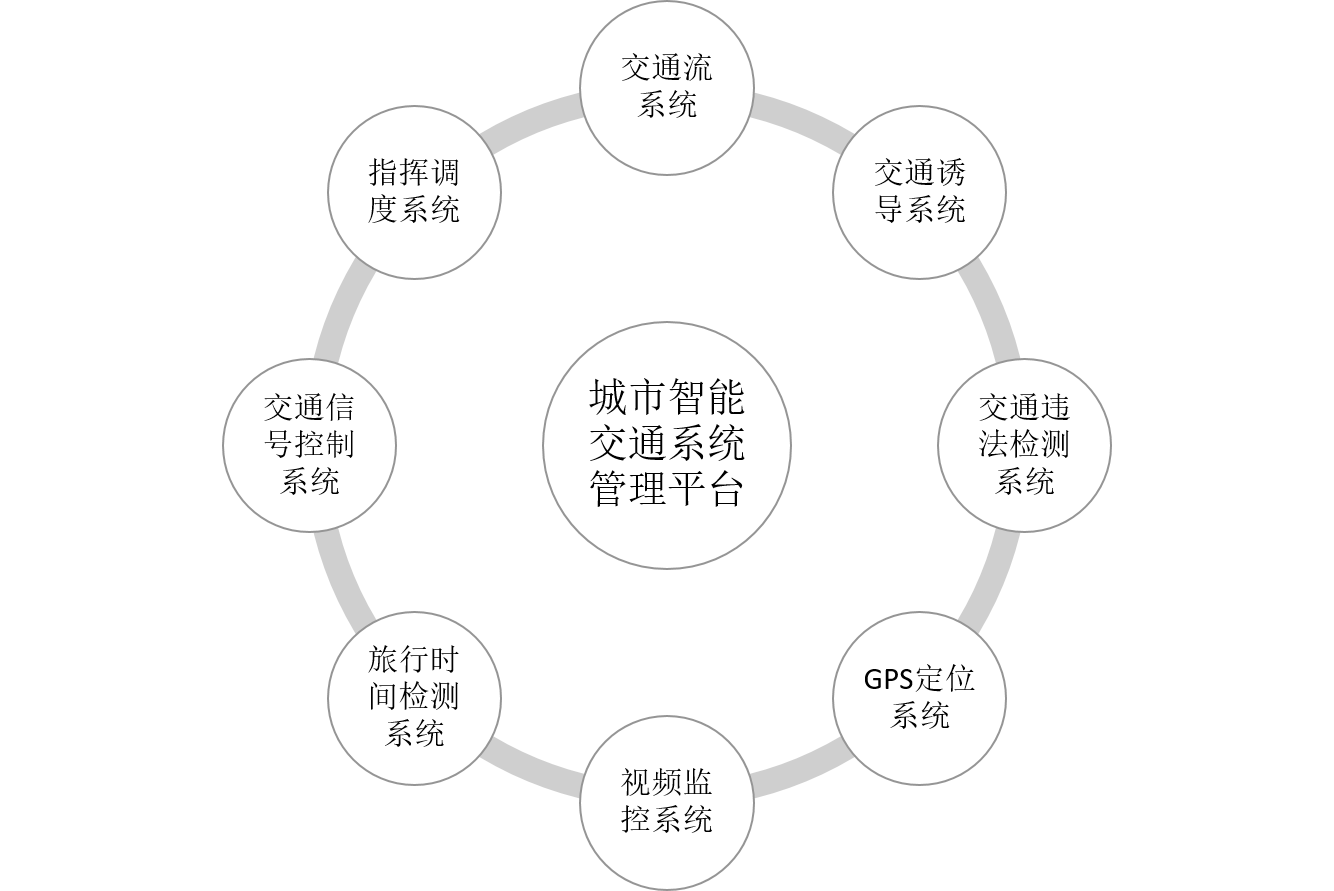


图1-2 智能交通系统概念图

由于传统的交通系统存在交通资源利用率不高的问题，导致车辆长时间的等待，而等待的过程不是那么长久，人们大多数情况下选择不熄火，造成燃油的浪费和更多车辆尾气的排放。

通俗地讲，智能交通系统就是根据实时的数据进行分析，全局把控，得到最优化的结果。其应用现状（2018年1月份），在北、上、广等高速发展的大城市中已经开始建设初的步智能交通系统；其中，北京在交通流的控制和导向、高速公路的管理和紧急事件的处理上应用了相关技术；广州主要在交通信息共享和静态交通系统改良上应用相关技术。

根据上图描述的智能交通系统，可将其大致分成三大部分：信息收集（视频监控系统、GPS定位系统、交通违法检测系统）、信息处理（交通诱导系统、交通流量系统、指挥调度系统）、信息应用（交通信号控制系统、旅行时间检测系统）。交通流量系统正处于信息处理中的核心地位，通过前者收集到的各种信息，经过交通流量系统的处理，得到实时的交通流，后续根据交通流进行相应的处理。因此交通流量系统的建立在一定程度上影响着ITS的效率，而仅仅一个交通流量系统的建立就很复杂。作为本科毕业设计，能处理交通流量里面的一个小的部分就已经足够。交通状况的实时监控是交通流量系统的主要任务之一。而对交通状态的把控，依据其特征参数有着不同的测量方式，目前普遍研究人员都是选择交通流量作为处理交通流的手段，通过对交通流量的预测到达对交通状况的预测。

交通流量预测，根据收集到的有关交通流量的相关因子，通过各种手段，进行数据分析得到交通流量与相关因子的某种关系，而交通流的复杂性使得其要到的一定的预测精度变得十分困难。

**1.2国内外研究现状**

交通流量的研究是研究交通状态的一种方式，同时也是智能交通系统研究的很重要的一部分。交通流量研究的最终目标是要达到对交通状况的定量描述和预测，构建精准的模型，通过输入的相关数据，在误差允许的范围内，对未来的交通流量进行精准的预测。对于相关数据的采集，一般采用的方法主要有以下几种：观察人员手动输入、车辆上的传感器、车辆通行电子信息卡、公路上的线圈检测器、雷达、摄像头等。随着大数据和硬件技术等的发展，采集相关数据的方式变得更加简便和迅速，在一定的程度上促进了人们对交通流量预测的相关研究。由于交通流量的动态改变，时间越短其预测的效果越好，在智能交通的体系下，短时交通流量预测的重要性更加凸显，因此目前都是是短时交通流量进行研究。短时交通流量预测，即对未来5至10分钟这个时间段进行预测分析。

目前短时交通流量预测相关的研究有很多，随着技术手段和数据采集的发展，相关工作达到的效果也越来越精确。交通流量预测即输入一个X（相关数据）得到一个Y（下一时刻的交通流），即一种回归问题。常用来解决回归问题的模型有：历史平均预测模型、时间序列模型、非参数回归模型、卡尔曼滤波、小波理论、突变理论、混沌、分形理论、人工神经网络、SVM、小波理论混合、神经网络混合等方法。

文献[3]等在Payne的MACK模型[4]上基于BP神经网络进行的高速公路动态交通流预测，使用MACK模型确定当前位置交通流量的相关因子，结合相关因子使用BP网络进行回归预测、文献[5]基于自适应指数平滑和径向基函数（RBF）神经网络的组合预测模型模型，使用结合一般方法和人工神经网络的方法，采取权值动态更新的策略，依据每个模型在前面的表现去调整这个权值，最后结合各种模型进行评估、文献[6]提出改进的BP神经网络进行流量预测,一方面动态修改网络的学习率，另一方面在学习的过程中增加动量项，使权值更新的过程中更加具有方向性、文献[7]提出一种将灰色关联度分析应用于实验前期的数据提取的卡尔曼滤波模型。使用灰色关联度进行时空相关性分析得出影响交通流的因素，再采用卡尔曼滤波利用提取出来的数据进行回归预测、文献[8]提出的基于粒子群优化神经网络的短时交通模型，使用粒子群算法对径向基（RBF）神经网络的参数进行随机优化，使得初始的参数在目前的情况下时最优的，从而能加快训练速度和模型的效果、文献[9]提出的基于遗传算法优化神经网络的短时交通模型，使用遗传算法优化BP神经网络的初始参数，达到类似文献[8]的效果，同时能在一定的程度上使训练的神经网络跳出局部的最优解、文献[10]提出一种使用改进的遗传算法优化BP神经网络的模型，使用模拟退火中的一部分修改遗传算法中的适应度选择问题，然后用改进的遗传算法对神经网络初始的权值进行优化，从而使BP网络训练的效果更好、文献[11]采用的在Spark下的改进的SA-SVR短时交通流预测研究，使用模拟退火（SA）对支持向量回归（SVR）的参数进行优化，防止其陷入局部最优解、过拟合等问题，从而在一定的程度上使SVR的训练得到改进、文献[12]采用多源数据对交通状态的预测，通过GSM切换进行定位采样得到车辆的速度等数据，将速度转换成空间平均速度的数据，通过SVR进行回归分析，得到回归模型、文献[13]使用核学习方法对短时交通流量进行预测，实验前对数据进行处理，使用核偏最小二乘的方法对数据中的影响因子部分进行提取，代入有效数据利用核极限学习机进行模型学习，之后利用核递推最小二乘加速学习速度、文献[14]提出的使用人工蜂群算法去优化神经网络的方法，类似如使用遗传算法或者粒子群算法一样，对网络模型的初始权值和阈值进行优质选择，加快训练速度和提高模型效果、文献[15]基于柯西加权最小二乘支持向量机的方式，采用遗传粒子群优化算法对其进行优化，从而更好的避免陷入局部最优解，以及加快到达最优解的过程、文献[16]基于ARMIMA模型的预测模型，采用矩阵估计、最小二乘以及BIC准测进行优化，使模型达到更好的效果、文献[17]提出基于小波变换的模型，将交通流量数据中的低频和高频区分开来。针对不同的部分使用相关算法进行回归预测，最后合成每个部分、文献[18]基于贝叶斯网络的组合模型，采用贝叶斯网络，将基于数学的模型的预测方法、基于统计模型的预测方法以及基于人工神经网络的预测方法的优势进行组合。

国外的研究比较早，早在1960年就有人提出AIMAM 模型[19]。随着不断的发展各种预测模型相继问世，传统的预测模型基本来之国外，根据现在的研究现状，人们开始从传统的模型转向人工神经网络。

文献[20]采取支持向量回归和多核学习的方法。文献[21]用遗传算法优化的神经网络去解决尝试。文献[22]提出了一种SIS流行病模型与复杂网络理论相结合的易受影响的复杂网络。

根据相关的参考文献和自己的查找情况，在处理交通流量的研究方法，主要方法由传统的数学模型，统计模型转向人工神经网络模型，同时各种模拟生物行为和进化的优化算法也逐渐加入。于此同时结合各种模型的集成算法也慢慢浮出水面。

**1.3 本文章节介绍**

第1章 概述。简单的介绍了一下研究背景和研究状况。

第2章 Tensorflow框架介绍。

第3章 短时交通流量预测的相关算法

第4章 短时交通流量预测系统设计。

第5章 短时交通流量预测系统实现。

第6章 本章对本论文完成的工作进行了总结。

**第2章 TensorFlow框架介绍**

**2.1 TensorFlow简介**

AlphaGo和李世石、柯洁的大战揭开了TensorFlow的神秘面纱，TensorFlow开始慢慢为人所知，它是由谷歌公司麾下的谷歌大脑团队开发的一种机器学习和深度学习的开源软件库，在很大程度上简化了机器学习和深度学习的过程。最初仅在该公司内部研究和使用，基于人工智能的重要性以及未来的发展，谷歌于2015年将其开源。

作为谷歌主要的机器学习和深度学习系统，TensorFlow可以运行在多个CPU和GPU上，通过多核、GPU甚至是TPU进行加速，从而减少训练的时间。与此同时TensorFlow不仅可以用在Windows、Linux、macOS等操作系统上，还能运行在移动端，包括Android系统和iOS系统。

TensorFlow将机器学习的整个过程，更确切的说将机器学习过程中的各种表达式中的数值和运算操作描绘成一张图，实际上它采用数据流图的方式进行数值计算，图中节点表示数学操作，图中的边则表示在这些节点之间传递的多维数据组（张量，可以理解成操作数）。图表示了该机器学习或深度学习模型的整个过程，图中保存着各个操作和张量之间的关系，这正是该框架简化传统构建神经网络的特别之处，神经网络最主要的问题就是训练参数，当神经网络的结构明确之后（对应该框架的图），之后就需要利用反向传播（BP）算法对其进行训练，即根据神经结构进行一次正向传播，即根据图的结构计算最终的值，然后将计算的值和真实之间的误差传递回来-误差的反向传播，根据其更新权值。正是由于这个原因，计算图能记住网络结构（运算过程），帮助我们进行正反向传递优化参数。对于一个表达式的计算，TensorFlow将整个过程划分为两个阶段，图的建立和图的执行。

构建一个用于计算2+3的图，代码如下：

01：import tensorflow as tf

02: # 创建一个变量a节点

03: a = tf.Variable(2, name="a")

04：# 创建一个变量b节点

05: b = tf.Variable(3, name="b")

06: #创建一个加法的节点

07: c = tf.add(a, b)

此时的c只是一个张量，没有具体的数据，只是构建了一个计算图。如果需要得到这个张量的值，我们需要执行这个图。就像是一个电路，线什么的都搭建好了，要想灯泡亮起来，我们需要打开开关。

执行这个图， 代码如下：

08: # 初始化变量

09: init = tf.initialize\_all\_variables()

10: # 启动默认图

11: sess = tf.Session()

12: # 运行图，三个节点a, b , add

13: result = sess.run(c)

14: # 输出结果6

15: print result

下面引用简书（[\_RayCloud\_](https://www.jianshu.com/u/49d1f3b7049e)）的一种理解[23]，可能更加容易理解，以下是上面计算全过程的一个计算图示意过程[23]。

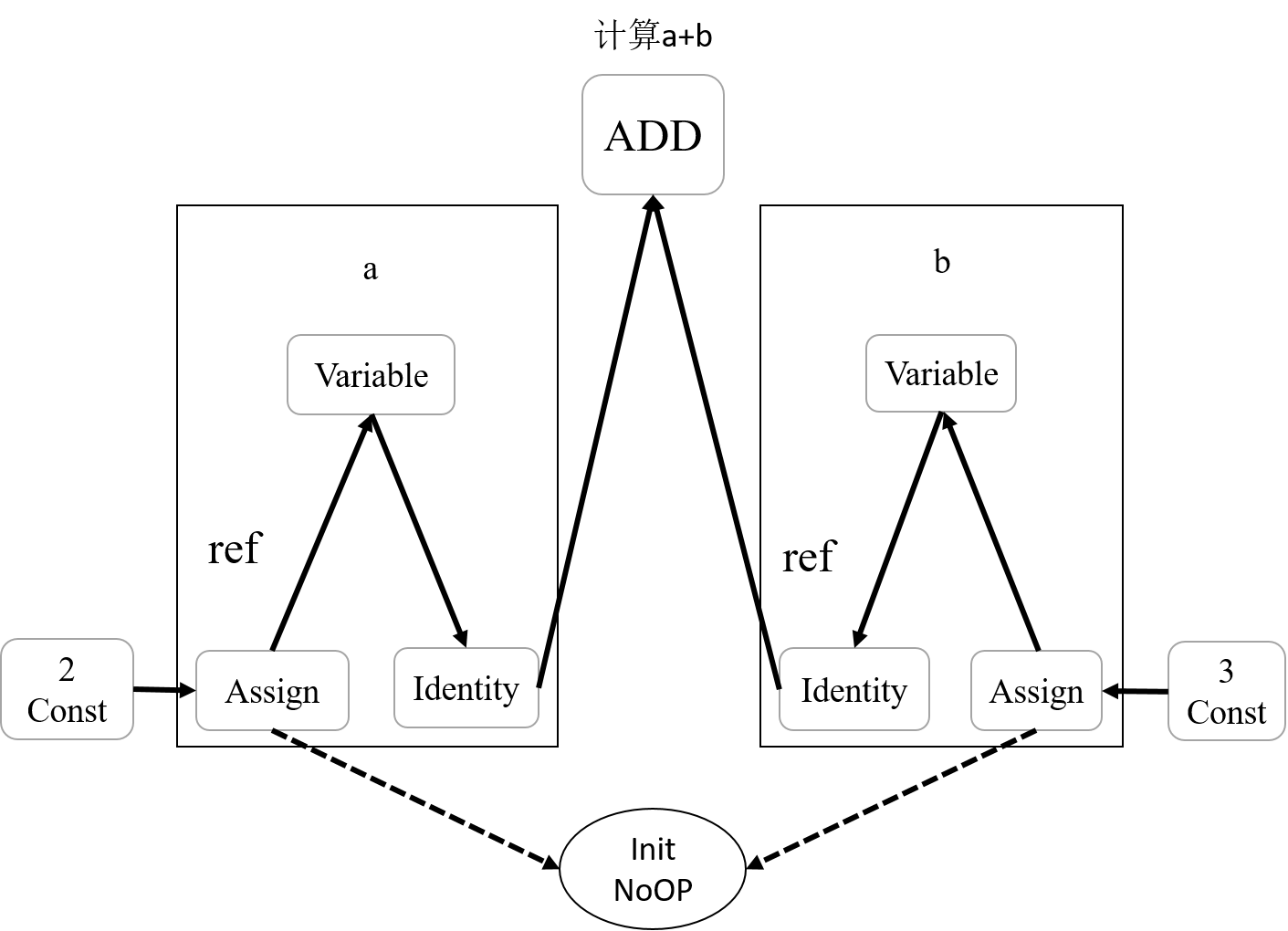


图2-1 构建的计算图模型

一开始，我们只是构建了a, b, c，以及常量2和常量3，此时还没有初始化，所有a,b只是张量节点，其中并没有值，只有进行Init操作才有数据，最后运行C，我们才开始计算三个op，得到最终结果。还有一点需要特别注意，就是图的概念，我们在使用该模型进行训练时，我们使用的是默认的图，在保存模型和打开模型时，特别是同时导入多个模型，我们需要使用不同的图，每个模型对应一个图的相关数据，如果在默认的图中打开所有的图，会发生参数冲突等问题。

**2.2 交通状态**

交通流是客观的数据表示，而交通状态是主观数据，随着不同的人在不断变化，虽然每个人对于当前状态的理解都有自己的看法，但都不外乎表达道路交通的拥堵程度，虽然每个人的感受不一样，但同样交通状态在人们的认识都落在一段差别不那么大的区间里，因此可以统计每个同样的交通状态（即客观交通流一样）对应的区间，这样就能交通流的测量值反过来表示交通状态。

交通流是指车辆或行人在道路上连续行驶或行走形成的流。一般使用宏观的交通流参数进行研究。相对微观的参数，宏观的更加具体。在研究交通流的状况时，我们本来就是粗略地研究交通状况的好坏，而不是去研究这条路上车与车之间的距离等。因此就放弃了对微观的交通流参数的研究，而是选择交通流量作为主要的研究对象，除了交通流量之外还有流速、密度、路段占有率等。

交通流量，根据传统定义就是在单位时间内经过交通道路上某一指定地点或断面的车量数。交通流量系统作为智能交通系统的一部分，我们在对其进行研究时，需要考虑交通流的状态，因此需要借用交通流理论定量地描述交通流。根据交通流理论，要定量描述交通流，我们需要以下三个参数：交通流量、交通流速度、交通流密度，而这三者之间的关系如下：

（2-1）

因此只要测量出交通流量就能很好地对交通流进行描述。可以通过对交通流量的研究，达到对交通流研究的目的。对于交通流量的测量，现在公路基本上都设有检测交通流量的设备，主要是环形感应线圈，在路面下埋设环形线圈，当有车辆通过，会引起磁场变化以此来计算交通流量。

**2.3 小结**

本章节，我们先简单的介绍了机器学习和深度学习库TensorFlow，即其利用计算图的概念来记住我们的网络结构，自动帮我们进行BP算法的过程达到简化机器学习或深度学习的过程，以及TensorFlow里面的几个常见概念，最后通过简单的实例程序进行演示其运行的过程。再讨论本文研究的主要问题，即交通流量预测。

**第3章 短时流量预测的相关算法**

机器学习类似如数据挖掘，从大量的历史数据中寻找潜在的关联或者规律。不仅仅如此，它还能学习到更深层次的内容，在这个数据中学习到的知识在一定的程度上能在新的数据集中运用并进行决策，并在一定的程度上能得到正确的结果，而交通流量预测问题，是一种回归问题，而我们需要解决的主要问题就是在历史的数据集中学习某种知识，并在未来的数据集中使用在这种知识并得到相关的结论。

机器学习，是一门很深厚的学科，它揽括多种学科，比如说计算机科学、工程学、统计学等。要想彻底弄明白需要有很深的数学功底，因此本章只是简单介绍一下机器学习中几个常见的可用于解决回归问题的算法。支持向量回归（SVR）、BP神经网络、长短期记忆（LSTM）网络，同时还介绍一种常见的优化算法-遗传算法。

**3.1 支持向量回归**

支持向量回归（Support Vector Regression，SVR），是一种处理回归问题的支持向量机（SVM）。传统的支持向量机（SVM）是用来处理分类问题，通过支持向量，简单将就是边界样本去确定一个超平面将多个类别区分开，对于当前维数无法解决的问题，可以通过核函数进行核映射到高维空间，再经过确定高维空间的超平面去解决分类问题。而支持向量回归（SVR），不是寻找超平面去区分类别，而是寻找一个所有样本都落在的平面去容纳它们。

支持向量机（SVM），在处理二分类问题时,其示意图[25]如需：

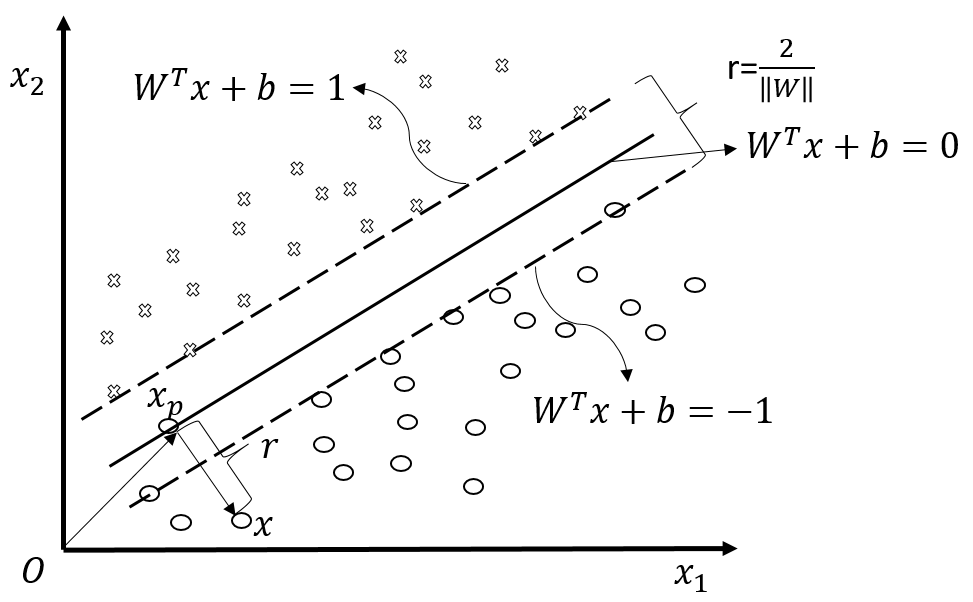


图3-1 SVM

通过中间的个由两条线组成的隔离带，为了得到更加精确的分类模型，因此我们需要这个隔离代最大化，假设任意一点X,X到分割线的垂直距离为r,在分割线上找一点,即有：

（3-1）

另有：

（3-2）

由于分割线有支持向量算出，=1,即分隔带为最大，得到这个隔离带之后，我们只需代入样本，看WTx+b的正负去决定其类别。为了得到这个隔离带，我们需要求的W，b这两个参数，由于通过样本进行训练，样本有特征（X）和标签（Y，类别），对于二分类问题Y（+1，-1），假设模型能正确进行分类，当将正样本代入表达式，其计算结果应该为+1，负样本代入进去，该表达式的结果应该为-1，总结这两种情况，可得，1。因此想要得到理想的模型，应该需要满足以下两个条件：

（3-3）

1 （3-4）

而支持向量回归（SVR），在处理回归问题时,其示意图[26]如下：

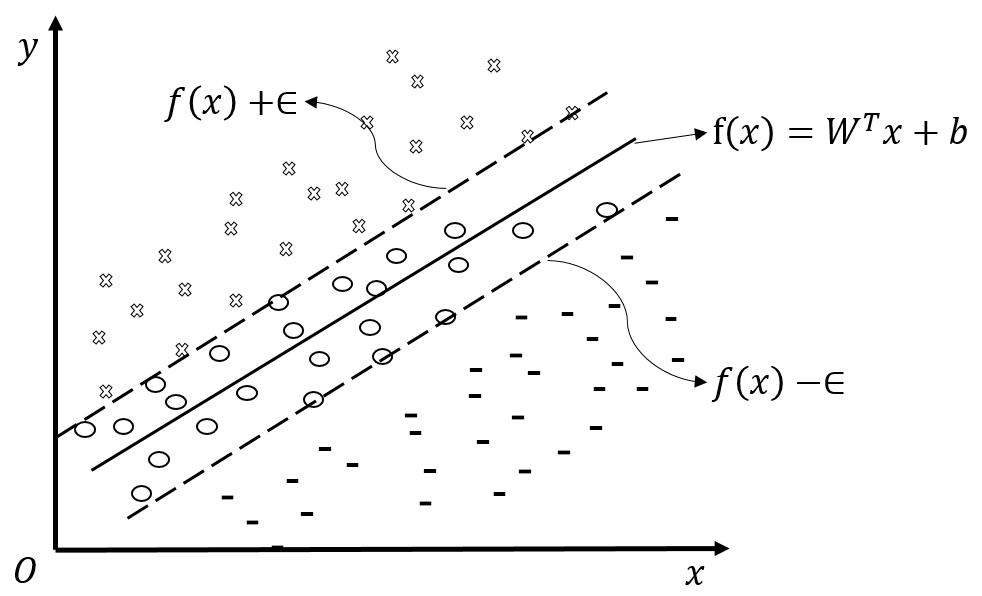


图3-2 SVR

我们需要的不是对不同样本进行分类，而是为了预测给定的样本特征，它代表的值时多少，拿最简单的线性回归而言，样本的X、Y满足某种函数关系，对于给定的X,就有与之对应的Y。而对于支持向量回归（SVR）而言，对于任意X，只要代入表达式f(x) = WT X + b,只要我的f(x)落在区间[f(x)-, f(x)+]就算正常，对于那些不在该区间内，我们就认为时错误。为了使模型更加准确，我们就需要减少这个错误，因此我们定义一个计算这个错误的计算式（损失）：

（3-5）

为了强调我们对的容忍度，我们需要定义一个新的参数，以及算计容忍的函数，因此上面的表达式化简如下：

（3-6）

根据我们上面讲到的，我们能容忍误差在的范围内的，因此-不敏感误差函数的定义如下：

（3-7）

通过引入松弛变量对表达式进行优化，对于简单的二维我们可以有如下便是，引用两个变量、,加入该参数后的示意图[26]如下:

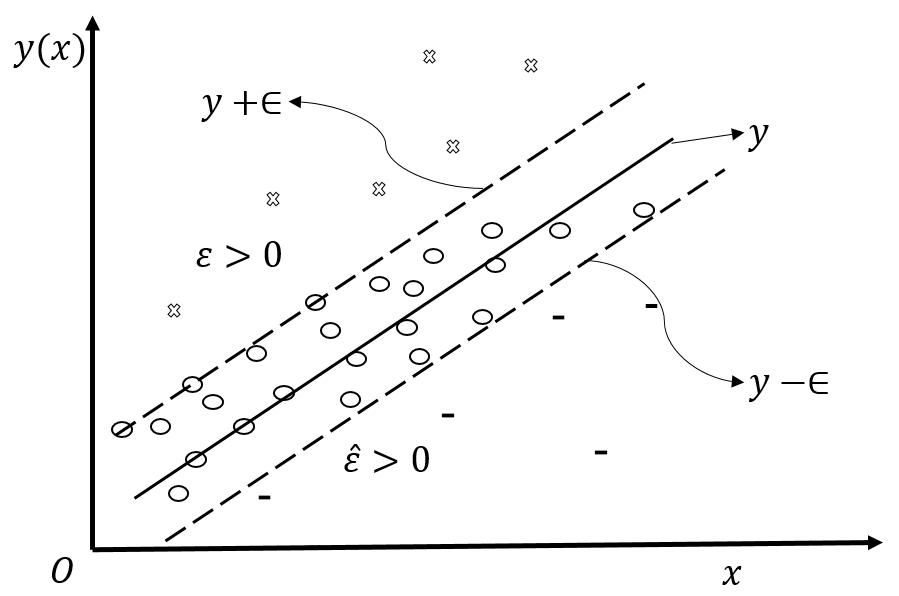


图3-3 引入松弛变量后的SVR

即在原始的间隔带两侧加入松弛带，如是损失表达式如下：

（3-8）

为了求出最终的模型，即我们求出参数W、b,我们需要满足一下的条件：

（3-9）

s.t. （3-10）

（3-11）

, （3-12）

然后根据拉格朗日乘子，引入对偶问题，根据KKT等进行求解。

**3.2 遗传算法**

对于有的问题，最优解是什么都可能无法知道，那么接近最优多半就没有什么规则可循，在这个情况下只能靠随机来处理，基于这个原因，美国密歇根大学约翰·荷兰教授提出了充斥随机因素的遗传方法（Genetic Algorithm）。

遗传算法[27]是一种借鉴生物进化的随机化搜索算法，在这个算法中，解被划为一代代的父、子、孙辈。每一代都是通过其父辈经过某种转化（模仿生物界的繁衍）而得到的。该算法的核心就是模仿生物不同代之间的进化和遗传过程。根据达尔文的《物种起源》，生物进化是优胜劣汰，适者生存，只有最适合“环境”的才能最容易活下来。由根据孟德尔遗传定理，生物繁衍的后代保留其父或者其母或者二者的基因。根据摩尔根的《遗传学说》，生物繁衍的后代又是由上一代基因突变或者组合，即有可能不包含其父或者其母的基因。因此在繁衍（进化）的过程中就包含：复制、交叉、变异。而达尔文的优胜劣汰就需要一个筛选的过程，即需要一个选择适者的函数，以及一个表现适应度的函数。以及现实中优化对象和基因型的对应关系（编码），基因和表现型之间的关系(解码)，以及表示解法的个体、种群。

综上所述，在一个遗传算法中，我们需要理解的概念就有：基因型、表现型、进化、适应度、选择、复制、交叉、变异、编码、解码、个体、种群。而其大体的思路，就是先根据现实的问题，找到一个编码方式，即每个个体如何表示解的形式，再决定一代中有多少个个体，即种群的大小，然后随机初始化种群中的每个个体，接着根据实际问题设计一个计算适应度的函数，即目标函数，一般为我们反映我们优化的因子的好坏程度，计算该种群中每个个体的适应度，再根据某种选择算法，选择这一代种群进行繁衍（交配、复制、变异等）产生下一代。直到产生的后代达到我们的预期要求。

遗传算法基本步骤如下：

（1）**解的表示（编码）**：再遗传算法中解被看着是独立的个体，对于个体是数的形式，我们能很好的进行比较、评估、选择，而对于非数值的形式，我们则需要根据具体问题的特点进行转化。但对于转换，我们首先得保证完备性，即具体问题所有解都能表示成数值，其次唯一性，即对于每一个解我们都有唯一的数与之对应。常见的解的空间的表示方法是二进制编码。

对于二进制编码，假设求解一个表达式在区间[-1,2]之的值。根据精度我们决定采取的位数，假设精度是0.1，则区间被分成30分，在，因此使用5位数进行编码，而此时的精度是3/32，11111表示2，00000表示-1，其解码就是：

（3-13）

（2）**初始种群的生成：**在遗传算法中，我们需要初始化第一代种群，因此我们需要确定种群的大小，即一开始，我们需要随机初始化多少个个体。

（3）**个体的评估：**为了进行进化，即所谓的“物竞天择 适者生存”，我们需要定义一个评价个体的适应度，而如何定义一个评价个体适应度的表达式，我们需要依情况而定，比如说求一个表达式的最大值，我们直接用该表达式进行评估，如果求最小值，我们就用表达式的倒数进行表示。

（4）**种群的自然选择：**竟然遗传算法是模仿生物的进化的，适者生存，就需要对种群进行选择，选择最优良的个体。一般采用轮盘赌算法或精英算法。精英算法，即选择种群中适应度最高的个体，而轮盘赌算法，则是根据每个个体的适应度占整个种群适应度的概率进行选择。假设一个种群4个个体的概率如下：

图3-4 轮盘赌算法

那么每一个个体的积累概率就是A：0.3、B：0.5、C：0.75、D：1，通过参数一个随机数，看它落在那个区域就选择谁。

（5）**种群的后代繁衍：**在选择了一个种群中的备选解之后，我们就要考虑如何生成新的解，保持种群的大小不变，一般通过复制、变异、交配、组合。

复制，即直接将父辈中的优秀个体放到到新种群中。变异，即在个体某一个位进行反转或者其他操作。交配，即两个个体部分结构加以替换重组生成新个体。组合，将上一代两个个体交配扩展到多个个体交配。

（6）**终止条件：**后代繁衍到什么样的状态才能算终结，由于“物竞天择 适者生存”的前提，随着繁衍的代数，个体也越来越满足人们对问题的需求。因此一般终止条件包括：最优解被找到、达到预期的条件、繁衍代数、更新无法得到更优解。

**3.3 BP神经网络**

神经网络是模仿生物神经网络结构，实现人脑部分功能的信息处理系统，随着McCulloch和Pitts于1943提出M-P神经元，紧接着Hebb学习，感知机，自适性线性神经元模型、最小均方差学习算法等人工神经网络相关的依次出现在世人的面前，这也导致人工神经网络的第一次高潮，1969年Marvin Minsky和Seymour Papert指出单层神经网络连最基本的异或也解决不了，而多层神经网络的学习算法还没找到突破口，致使人们开始转向其他方向。峰回路转1986年Rumelhart提出反向传播(BP)算法，解决了神经网络学习的复杂问题，为神经网络带来了新的生机，加之传统人工智能算法在视觉、声音等方向的不足，人们又开始转向神经网络，随着神经网络在图形处理等方面取得成功，神经网络又重新回到了人们的视野中，并且在社会上掀起了学习神经网络的大潮。

生物神经网络结构图[28]如下：

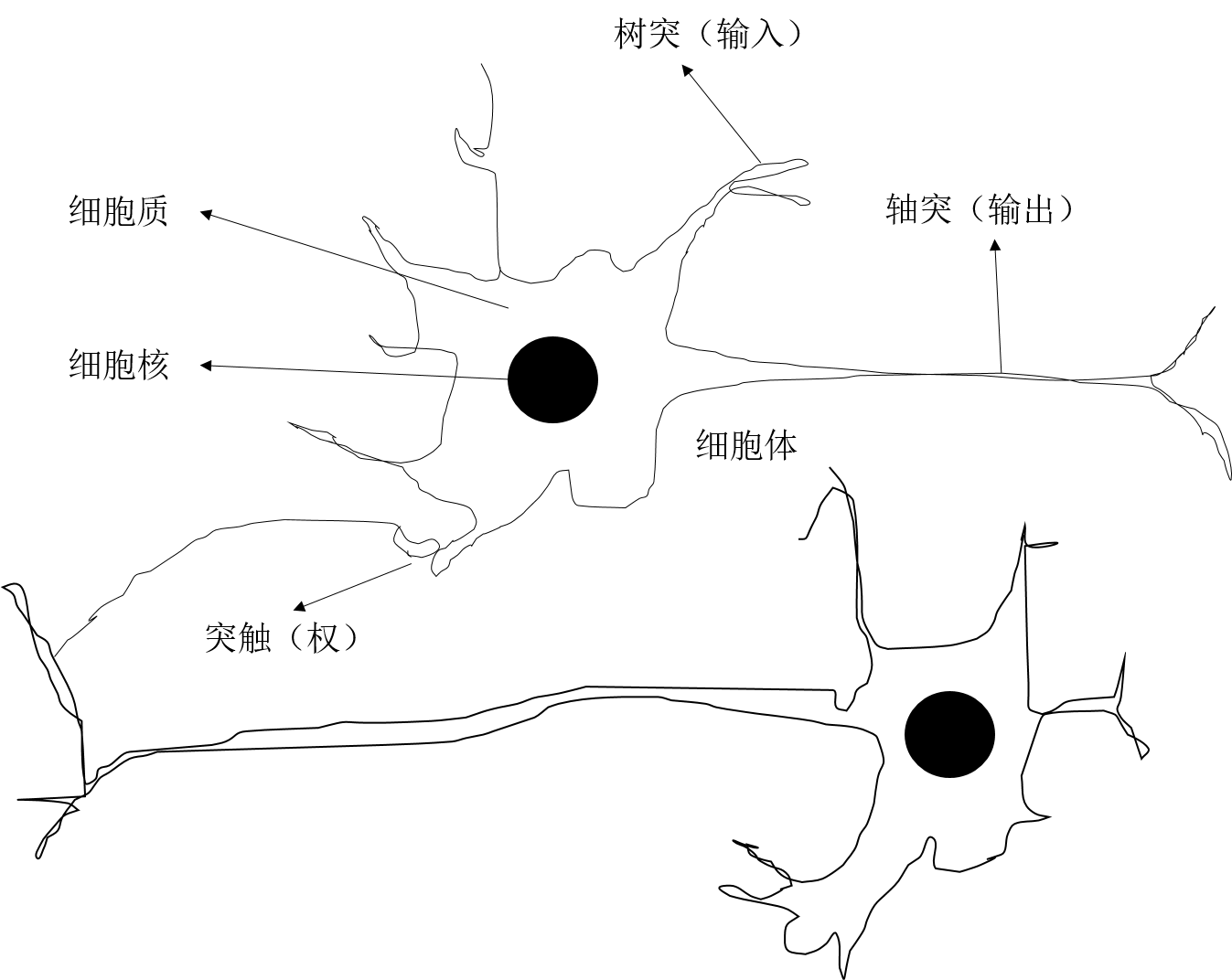


图3-5 生物神经网络

一个细胞树突接受其他神经元的突触传递的刺激，根据细胞体经过相关处理，决定是否向轴突发送信号，通过学习生物神经细胞的结构，人们模拟的人工神经网络示意图[29]如下:

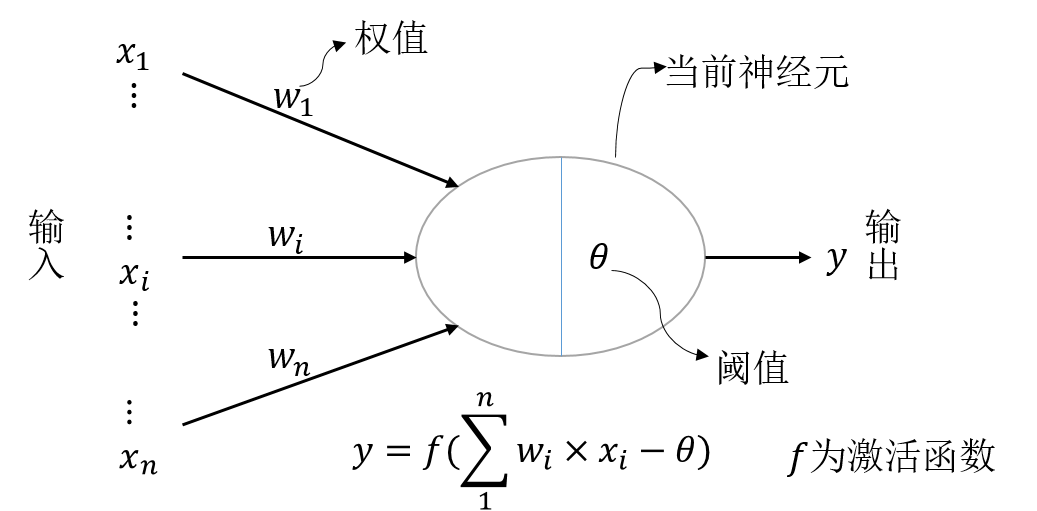


图3-6 M-P神经模型

当前神经元接受来自其他神经元的的输入，依据权值进行计算得到结果，在根据“激活函数”决定是否向下传递“刺激”。

一般我们称之为BP神经网络的模型是使用BP算法进行学习的多层前馈神经网络，如下是BP网络的模型示意图[30]：

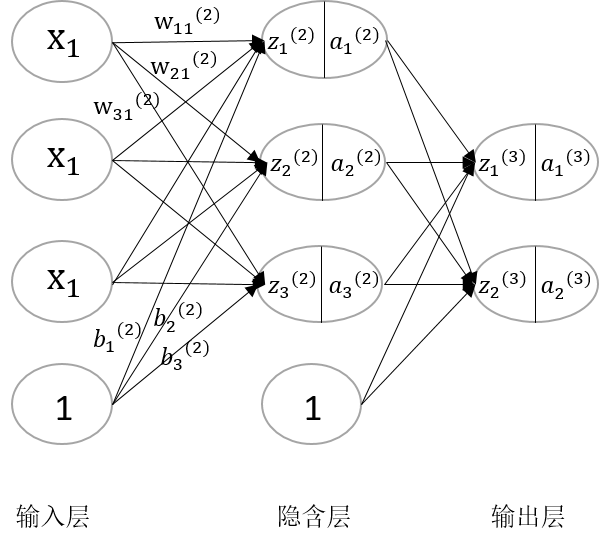


图3-7 BP神经网络模型

根据模型的关键，我们可以计算输出结果，一般最后输出层不需要进行“激活函数”的操作，一般“激活函数”有relu,sigmod等。

（3-14）

（3-15）

（3-16）

为了通过样本进行训练得到我们想要的结果。类似支持向量回归（SVR），我们定义一个损失函数进行评估模型的好坏，为了得到我们想要的模型，损失函数的值应该尽量小，我们定义最小方差损失函数：

（3-17）

为了得到所有的参数W、b,我们需要在每轮更改参数W、b使得loss变小直到达到我们的需求,而如何进行更新，我们使用BP算法对权值W和偏置b进行更新，BP算法基于梯度下降（gradient descent）,每次向负梯度的方向移动,而每次移动多少，取决于学习率的大小，参数更新表达式如下示：

（3-18）

（3-19）

根据高等数学中的知识，在求导时，对于间接求导，我们通过链式法则，得到如下的表达式：

（3-20）

假设我们计算了所有权值W和偏置b的偏导数，那么BP网络的算法如下：

1、选择激活函数和学习率

1. 初始化W、b，随机。
2. 代入样本，计算预测值
3. 计算W、b的关于loss的偏导数
4. 进行更新
5. 如果没有达到预测设定的结果，重复3，4，5。

一般神经网络训练容易陷入局部最优解，因此在上一小节介绍了，优化算法遗传算法，虽然遗传算法不能解决回归问题，但使用遗传算法对W、b进行初始化，在一定的程度上能降低得到局部最优解的概率，同时也能加快神经网络的学习速度。

**3.4 长短期记忆网络**

随着神经网络的出现，传统方法在语音、翻译等工作上的缺陷，人们开始选择采用神经网络进行处理，催生了循环神经网络（RNN）的发展，循环神经网络的关键点之一就是可以用来连接t-1的时间步到时间步t上，即RNN在第t-1个时间步的判断对第t个时间步的判断会造成影响。

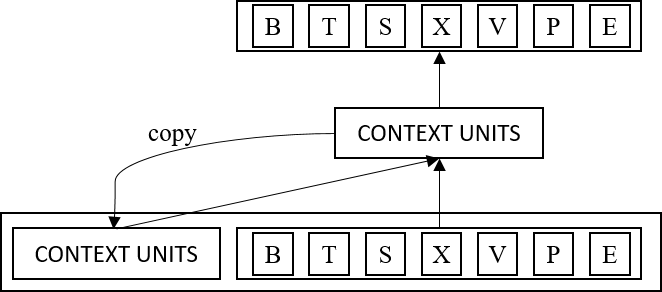


图3-8 Elman早期RNN的示意图

上图[29]中，BTSXVPE 代表输入，而 CONTEXT UNITS 则表示前一时刻的输出。将前后时间上的数据结合起来，即使前一时间点的数据影响后面的输出。

借鉴Christopher Olah[32]的《Understanding LSTM Networks》中提到，假设我们有一个基于前面的词推测现在位置的词的系统。对于“the clouds are in the sky”最后一个词，我们不需要上下文就可以猜出来，而对于“I grew up in France... I speak fluent French”，对于最后一个词，我们如果不依赖很早之间的信息我们就只能推测出是一种语言，而该语言是何种语言，我们无从得知。而如果我们想弄清楚是什么语言，我们就得记住很远的上下文中的关键信息，相关的信息和当前预测的位置之间的间隔很大。

一般的神经网络都是通过BP（Backward propagation）进行训练，计算W，b关于最终损失函数的偏导数，通过梯度下降进行调整的。然而依赖的信息和预测的位置存在间隔很大的问题，链式法则我们知道，由于距离的加长，我们最终求得的偏导数有可能消失或者膨胀，即所谓的梯度消失问题，也是RNN的长期依赖问题。

长短期记忆（Long Short Term Memory）网络是一种改进的RNN网络，能够解决RNN中的长期依赖问题。LSTM 由Hochreiter & Schmidhuber 提出，之后又很多人对其改进。下面是其中一个流行的 LSTM 变体，由 Gers & Schmidhuber 提出的，增加了窥视孔连接(图中的虚线)，将当前记忆单元Cell的状态传递给门控单元。

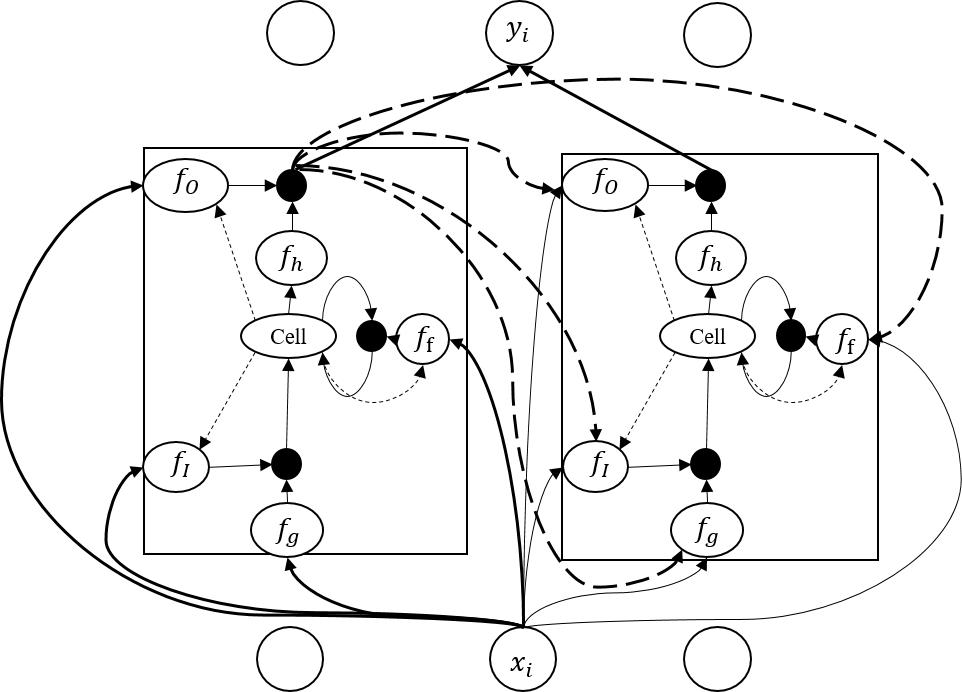


图3-9 LSTM网络

由上图[33]可知，相比传统的神经网络，传统神经网络中的神经元替换成了黑盒子Block，正是这个Block解决RNN中的长期依赖问题。Block细节图[33]如下：

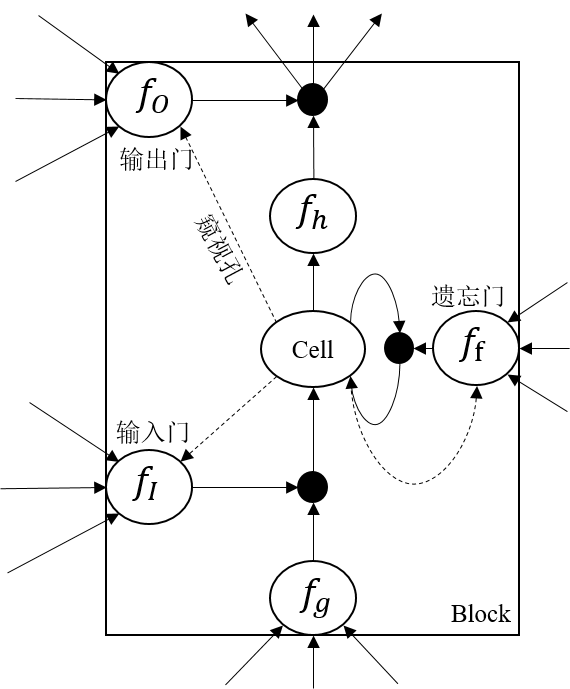


图3-10 Block单元

每个block包含一个或多个Cell和三个门控单元Input Gate、Forget Gate、Output Gate。三个门是非线性求和单元，经过相应激活函数其值为（0/1），通过相乘点控制相应的输入是否成功。输入门通过输入收集的激活信息和输入门的状态控制Cell的输入，相乘点通过输出门的状态控制Block的输出，而忘记门决定是否保留Cell的先前状态。cell内没有激活功能。从Cell到各种门连接（上图中的虚线,这也是一种改进的LTSM），表示将Cell状态乘上一个权值（每个门的权值独立）传递给三个门口单元。块内的所有其他连接都未加权。从块到网络其余部分的唯一输出是Block最终的结果，即块输出，同时块输出的个数取决于cell的个数，共用一个output gate。

综上所述，LSTM的基本单元Blck包含的内容：

3个控制门，input gate、forget gate、output gate分别控制输入，上一个状态以及输出。

1组存储器，包含一个或多个记忆单元，其值表示cell state。

cell收集器和输出器，前者收集并计算准备传给cell的值，后者cell的值激活后准备作为块的输出。

若干个相乘点。

block output，该值传递给其它 block（也包括自己）的门控单元，以及输出层或下一个层次。

上面是LSTM的基本知识，下面根据Christopher Olah的《Understanding LSTM Networks》进一步对LSTM的内部工作进行详细的介绍。

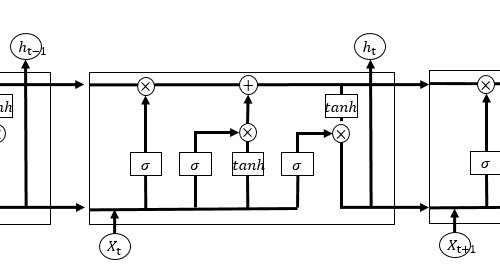


图3-11 LSTM整体模型

上图[32]从左至右依次是遗忘门、输入门、输入、输出门。

对于遗忘门：

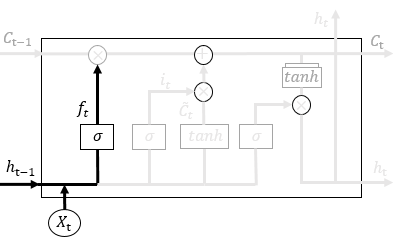


图3-12 LSTM中的遗忘门

（3-21）

上图[32]介绍的是遗忘门，式3-21表示计算过程，权值和上一个输出、当前时刻输入以及偏置，再通过某种激活函数σ作为遗忘门的值（0/1），控制上一个状态是否被保留。

对于输入：

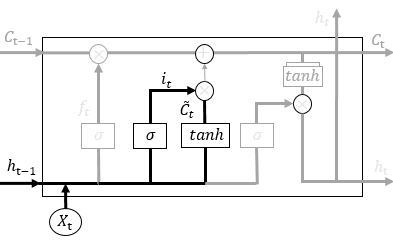


图3-13 LSTM中的输入门和输入

（3-22）

（3-23）

上图[32]表示输入门，式（3-22）是计算过程，是由权值，以及上一个输出，当前时刻的输入，以及偏置，通过某种激活函数作为输入门的值（0/1）来控制输入是否能进入，表达式（3-23）表示输入，是由权值，一起上一个输出，当前时刻的输入，以及偏置，通过激活函数tanh(双曲正切)作为输入的值。

更新cell的状态，由遗忘门结合cell前一时间的状态，以及经过输入门修正的输入来更新：

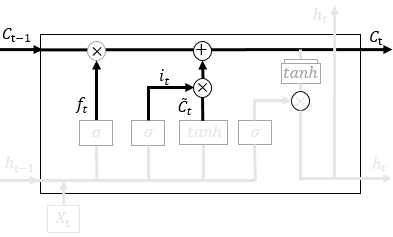


图3-14 LSTM状态更新

（3-24）

上图[32]表示Cell更新过程，表达式（3-22）表示更新Cell中的状态，其由式3-19的遗忘门，控制前一个状态，以及表达式（3-20）输入门，控制输入的状态。

总上所述，我们已经介绍完LSTM的基本单元black，以及其组成“元件”和里面的运行过程。对于带有窥视孔的LSTM，我们只需要将输入增加一项状态而已，如下图[32]示：

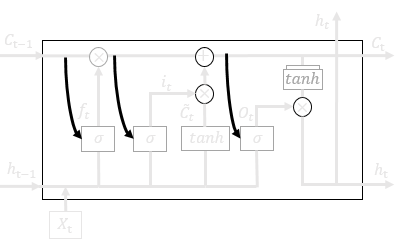


图3-15 带有窥视孔的LSTM

之前的公式里面的两项变成三项，如下对于输入门而言（其他的类似）：

（3-25）

（3-26）

表达式（3-25）转化到表达式（3-26），对于每一个门而言增加考虑Cell前一个状态的信息。

在此我们已经介绍了LSTM的基本组成单元，以及其内部的工作流程。实际上其和普通的循环神经网络仅增加了一个block单元，解决长期依赖问题（梯度爆炸或梯度消失问题）。因此在使用的过程中，我们不仅需要设计神经网络还需要设置相关的Cell。

**3.5 小结**

本章主要简单介绍几种处理回归问题的方法，在理论上说明解决该问题的方法的多样性。为系统的建立提供理论基础。从支持向量回归，到BP神经网络，再到循环神经网络。每一种方法要深入理解的话，需要了解很多相关知识，由于时间和能力有限，只能简单地介绍一下。

**第4章 基于多源数据的高速公路流量预测系统设计**

我国高速公路的不断发展，车辆行驶在高速公路上变得很常见，因此也让高速公路的维护变得越来越困难，其主要难点：一、高速公路的拥堵问题。二、对车辆的及时救援。三、及时查找高速公路损坏部分。虽然看起来没有什么关联，但我们都能根据交通流量进行分析，如有时候，交通流量达不到预期的预测值，我们就怀疑其可能发生交通拥堵，或者发生事故。对比历史个检测点的数据，我们甚至可以找出测量点之间是否由于其他因素导致数据不一致，假设某两个测量点之间的路段出现细微的损坏，则在一定的程度上影响测量点前后的流量的轨迹。因此对交通流量进行很好的预测的话，那么在一定的程度上能解决这些问题。

基于多源数据的交通流量预测，根据多种与当前交通流量相关的因子作为输入，得到我们要预测的交通流量。在本系统中，我们选择的相关因子，主要使历史交通流量数据和天气因素。

**4.1 需求分析**

根据给定的G2（京沪）和G42（沪蓉）的测量点数据，以及周边天气因素，在误差尽可能低的情况下，对交通流量实现短期预测，即预测5至15分钟（3个数据点），由于测量点的数据格式一致，周边环境一直，因此我们只需要做一个对于14个测量桩通用的软件就行，即我们可以选择数据集进行训练，得到该数据集上的模型，然后可以在验证集上测试以下模型的效果，再决定如何进行参数调节，等模型训练完成后，再可以对未知的交通流量进行预测。

**4.2 总体设计**

**4.2.1 系统结构设计**

1、处理样本的结构，即第二章中对样本进行提取的功能模块。

2、各种预测模型的模块，即使用不同的预测算法在数据集上进行训练，得到不同的预测模型，该模块需要进行多次尝试，使模型达到我们预期的要求。

3、预测模块，即将当前的数据导入训练好的模型中，依据各个模型的权值，对每个模型得到的值进行加权求和，得到最后的预测值。

**4.2.2 系统流程设计**

对于整个系统而言，我们第一步就是处理数据，即从原始的数据到我们需要的训练样本，就是机器学习中常见的特征、标签对，其次就是我们的训练过程。使用机器学习的方法进行回归预测，训练过程就是修改模型参数的过程，调节参数使得我们的模型吻合度变好，因此训练过程事关重要。同时在预测之前，我们需要通过测试我们训练的样本是否符合我们的要求，使用测试集中的数据在训练好的模型上进行测试，根据测试的结果，考虑如何调试我们的模型参数，模型决定后，我们就可以根据当前的数据进行预测，得到未来的交通流量。

提取样本的流程图（参照第二章），从原始数据中抽出我们需要的数据进行训练。

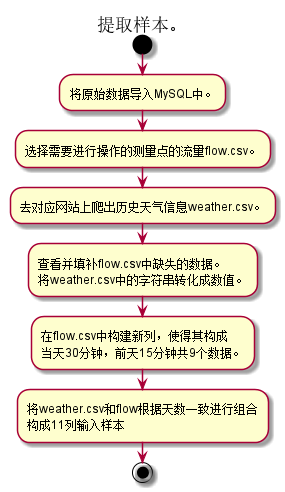


图4-1 提取数据样本的流程图

对机器学习算法的训练过程，将提取的数据集进行分类，分成训练集和测试集，然后将多种机器学习的算法在训练集上训练，在测试集上进行测试，根据测试结果进行调参，最后组合几种表现良好的算法模型进行集成回归：

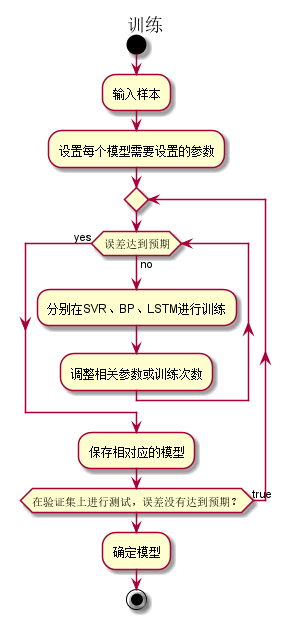


图4-2 对数据进行训练的流程图

对训练好的模型进行预测的流程过程：

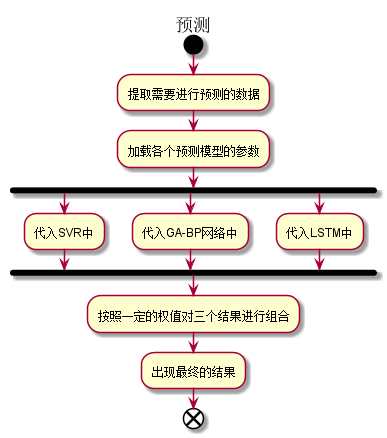


图4-3 使用模型进行预测的流程图

**4.3详细设计**

**4.3.1 处理输入数据**

由前面第二章介绍的，我们的输入样本是当天半个小时6个数据，当天的前一天的十五分钟3个数据，以及天气2个数据，以及需要预测的5至15分钟的3个数据，最后加上时间共15列的逗号分隔符文件（.csv格式），我们使用Python中的文件处理函数open对我们的数据进行读取，依次处理每一行的数据，根据逗号进行分隔，成15列，我们把前2-4列，即我们需要预测的数据作为标签（Y），我们后面11列，即所谓的历史交通流量，以及当时的天气因素作为特征(X)，为了使回归模型的效果更佳，我们需要将整个数据集划分为训练集和测试集。训练集是让我们的模型学习到一般的规律或知识，而测试集是用来验证模型的泛化能力和进行参数调节的指向标。

参考周志华老师的《机器学习》，传统划分数据集的方法有以下几种：留出法、交叉验证法、自助法。留出法：直接将数据集的一部份拿出来作为训练集，另一部份的数据作为测试集，一般都是将训练集和测试集进行七三分。交叉验证法：先将数据分成K等份，每次拿出k-1份作为训练集，一份作为测试集，一共进行K此，最终取均值。自助法：假设数据集D有n的样本，我们对D进行有放回抽样进行n次形成训练集，根据概率论知识，训练集中除去重复的数据之后大约占数据集中的70%。

本次我们使用的留出法，七三分。同时为了加快训练的速度和模型的精确度，我们使用下列公式。同时打乱样本的顺序，使其减小算法模型学习无用规律的概率。

（4-1）

式4-1是零-均值规范化，x减去x的均值，除以其方差，将其转化成标准正太分布。

**4.3.2 定义模型参数**

在这个系统中，我们用来进行预测的模型有三种，分别是支持向量回归、GA优化BP神经网络、长短期记忆模型。

对于支持向量回归而言，我们设定的参数容忍度C、松弛因子以及核函数。对于GA-BP我们需要设定的参数有GA里面的种群大小，个体的编码方式，遗传的代数，变异的概率，交叉的概率。对于BP神经网络，我们需要设置神经网络的容量，即该神经网络的层数和每层的神经元个数，模型学习的速度，学习算法的选择。对于LSTM，除了传统的类似BP神经网络相关的参数，我们还需要确定序列的大小，Block的结构。

**4.3.3 确定模型评估方式**

传统的模型评估一般是计算预测值和真实值之间的差距，一般而言就是使用误差百分比，即用真实值减去误差值的绝对值再除以真实值。解释方差(explained\_variance\_score)、平均绝对误差（mean\_absolute\_error）、平方误差均值（mean\_squared\_error）、R2\_score。在这里我是直接使用sklearn(一个机器学库)里面的回归器性能评估方法。

explained\_variance\_score，假定是预测值集合，y是样本的实际值集合，Var是方差，那么解释方差的表达式如下示：

（4-2）

解释方差的最大值是1.0，表达式的值越高，模型的拟合程度越好。

mean\_absolute\_error，假设是第i个样本的预测值，是实际值，一共有个样本，那么该方法使用的表达式如下：

（4-3）

由表达式我们可以知道，当该表达式的值越大其效果越不好，同时还存在一个问题，由于该表达式是直接使用预测值和真实值相减，则会根据量纲的不同而不同。因此再使用该方法进行模型评估时，我们需要将数据进行归一化。

mean\_squared\_error，如上，假设是第i个样本的预测值，是实际值，一共有个样本，那么该方法使用的表达式如下：

（4-4）

类似平均绝对误差，该表达式的结果越小，代表该模型的效果越好，同上，该方法也存在量纲问题。

R2，如上假设，是第i个样本的预测值，是实际值，一共有个样本。

那么该方法的计算过程如下示：

（4-5）

由表达式而言，该表达式的值越接近1，代表其模型的效果越好。

误差百分比，假设是第i个样本的预测值，是实际值，一共有个样本。其表达式如下：

（4-6）

error的值越小，代表其整体效果越好。

**4.3.4 界面设计**

对于这个系统，我们需要**选择训练数据集**，选择完训练的数据集之后我们需要对各种回归算法进行**训练**，训练完成后，我们需要直观地**查看**我们训练的结果，除此之外，我们还需要了解我们模型的泛化能力，即我们需要在验证集上进行**验证**，最后，当我们的模型训练完成之后，我们就能开始进行**预测**，因此，界面初始要求就是，4个按钮（选择、训练、验证、预测），一个显示图片的（显示训练误差、验证误差），一个三个用来显示预测结果（5、10、15分钟的预测值）。

**4.4 小结**

本章主要对基于多源数据的高速公路流量预测系统进行详细设计，介绍系统的结构模型，以及相关用来评估实验效果的方式。

**第5章 基于多源数据的高速公路流量预测系统实现**

**5.1 系统开发环境**

* 操作系统：Ubuntu16.04。
* 开发工具：Pycharm。
* 使用到的第三方库：TensorFlow 1.0、sklearn、pandas、numpy、matplotlib等。
* 编程语言：Python3.6。

**5.2样本提取**

本节以G2京沪高速、G42沪蓉高速，南京到上海段，即从南京沿着G42到无锡惠山区与G2相交，再从交点到上海。

**5.2.1 数据准备**

本文的交通数据来自于南京市交通局，一共14个测量桩，记录着往返8个车道上的车类型，车速，车距，流量。时间间隔为5分钟。流量，即5分钟内经过测量桩的车数目。记录时长从2017-06至2017-08共三个月。由于研究的课题是基于多源数据的高速公路流量预测，因此除了流量，还有天气、上下游相关收费站数据。对于与天气相关的数据，我们选择去相关测量桩附件的气象局网站上去查询。对于与收费站相关的数据，来自于交通局。（最后发现收费站数据时间实际与交通流量不匹配，故舍弃）

**5.2.2 工具介绍**

对于提取流量需要在大量的数据中进行车类型合并，车道合并等，该数据大约有73万条，且该文件为.xlsx，目前有很多编程语言都支持对其的操作，而数据库对于大数据处理有着天然的优势，所以选择导入数据库，在此本文使用了免费的数据库MySQL。对于天气数据，由于其来自于Web网站，因此可以通过编写一个简单爬虫进行爬取，在这里我们使用现在比较流行的bs4库和用于动态加载网页的selenium工具（由于该Web网站上的天气数据是动态加载的）

**5.2.3 提取步骤**

1、提取数据流信息。

原始数据一共有34列，包括：录入时间（录入数据库）、记录时间、车道、每种类型车的数目和速度、桩号等。如下示：

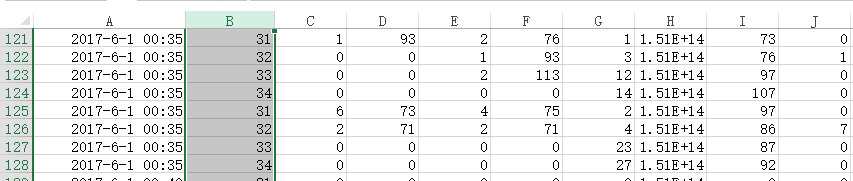


图5-1 原始的Excel数据（部分）

Step1:将Excel中的数据删除没用的列，再另存为.csv格式，在数据库中创建相应的数据库以及表，其中表的属性列和该.csv文件里面的列要一一对应。

Step2：使用SQL语句将.csv导入数据库中，再根据我们需要的信息，进行相关的SQL语句导出自己需要的数据。

2、搜集天气数据

查询到测量点的地点，去相关的气象局网站寻找数据。在这里我们使用的“天气后报”本来打算去中国气象局，发现精确到小时的数据，只能下载一周的历史数据。而自己想要免费获取历史的每小时的天气，在寻找了很久之后，只有收集历史数据以天为单位（免费情况下）。

用数据集里面测量桩编号为K1167\_470（距离G2起点1167千米470米）的测量点，根据查找显示其坐标大致在昆山附近，因此去“天气后报”网站中，我们查询到昆山的历史天气，由于我们的车流量数据是2017年6、7、8三个月的数据，因此我们需要对应相应的时间，如下是2017年6、7、8月的历史天气的网址：

<http://www.tianqihoubao.com/lishi/kunshan/month/201706.html>

<http://www.tianqihoubao.com/lishi/kunshan/month/201707.html>

http://www.tianqihoubao.com/lishi/kunshan/month/201708.html

我们打开6月份的网址显示如下：



图5-2 网页布局

用Google浏览器打开开发者F12，我们可以查看到，其主要内容显示在一个类选择器为“wdetail”的div下面的table下面的tbody，即找到我们需要的数据在HTML界面中对应的标签。

Step1:下载我们需要的界面。找到对应的标签。

Step2:使用相关库提取标签中的内容。

在实验过程中，开始进行简单的尝试，发现在对应的标签中找不到需要的数据，在同学的帮助下，发现该数据是动态加载的。因此使用库selenium，相关使用步骤可以参考官网，提取好的数据进行相应处理，得到如下结果：



图5-3 处理后的结果

**5.2.4 提取步骤**

现在，我们已经有了2017年6、7、8月份某一个检测桩测量的交通流量数据、该检测点附近的收费站的车量进出数据以及该检测点附近的天气数据。根据MACK模型，我们应该采集路段上下游数据，但对于高速公路而已，我们可以采用测量点上下测量点代替，但由于我们获得的测量点之间间距比较大，因此我参照浙江省交通科学研究院[24]，选择该检测点前一天前15分钟的数据3个，以及该测量点当天前30分钟的数据6个，以及天气和风速共2个。

如下显示，是我们处理后的交通流量和天气因素的数据：

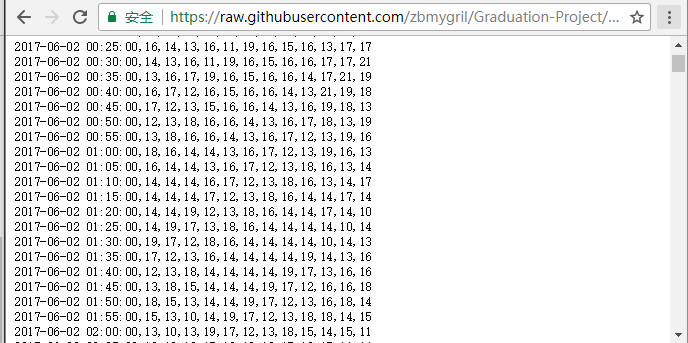


图5-4 交通流数据



图5-5 数字化后的天气

由于交通流量数据在采样时存在缺失，因此通过简单的处理，对于非节假日，我们通过环比，采用一周前的数据，或者一周后的数据进行填充，对于填充后的数据，如果出现较大误差的，进行手动处理，最后将交通流量数据和天气数据根据日期进行合并。

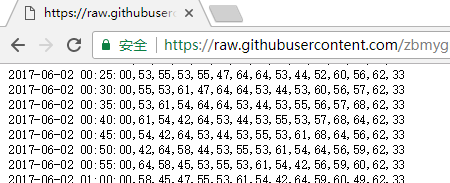


图5-6 最终的样本格式

**5.3 系统实现**

**5.3.1 输入数据功能实现**

对于数据的输入，由于在这个系统中我使用了TensorFlow的框架，因此在这里我尝试了两个途径，途径一：一般的文件读入，即将数据通过文件读入的方式读到内存中的数组中存储，在使用的时候，用TensorFlow，feed\_dic方式传递给图中的placeholder节点。途径二：使用TensorFlow的读取文件的方式，让TF自己从文件中读取数据，避免读入内存，再Feeding的过程，减少开销，一般适合读取数据量大的数据。同时其可以设置是否打乱顺序，设置bitch的大小，以及整个样本的循环次数。

**5.3.2 模型训练功能实现**

在该模块中，我们使用的又是存在的库，而不是重新造轮子，对于具体的代码可以观看我的GitHub地址<https://github.com/zbmygril/Graduation-Project>，而此处只是简单的介绍一些参数设定的问题。

对于支持向量回归（SVR），我们首先设置SVR的容忍度C和松弛因子，，然后设置核函数，将低维空间向高维空间转化，对于非线性回归，我们一般使用的径向基函数（RBF），然后就是利用拉格朗日乘子，对偶问题等进行解决。如果全部自己实现的话，会比较复杂，由于现在有大量的机器学习库中包含该模型，对于SVR我选择了常见的Sklearn库，调用的sklearn机器学习库中的支持向量回归，对于库的使用，我们只需要在调用时设置参数：容忍度、松弛因子、核函数。再就是将我们的训练集数据传给SVR模型作为参数即可。而在这个系统中我使用的参数C=1.0，松弛因子为0.5，核函数为RBF。

使用支持向量回归的伪代码如下：

-------------------------------------------------------------------

算法：支持向量回归

-------------------------------------------------------------------

输入：带标签的训练数据集，测试数据集

-------------------------------------------------------------------

输出：训练好的模型

-------------------------------------------------------------------

01：导入使用的库函数sklearn

02：定义松弛因子、容忍度、惩罚力度C、以及核函数

03：调用库中的函数，设置我们定义的参数，生成支持向量回归类

04：调用支持向量回归类中的训练函数，设置数据集为训练数据集

05：保存训练好的模型

06：使用模型，代入测试数据集，调用该类中的预测函数

07：得到测试数据集对应的预测数据

-------------------------------------------------------------------

对于BP神经网络。首先设置网络的结构。主要是神经容量大小的设置，更朴素的说法即神经网络的层次和每层的神经元个数，其次是学习率、训练次数。对于传统的做法，先进行一次前向，即每个数据得到的结果，再来一次后向，对权值进行更新，对于后向我们需要对每个变量关于损失函数的偏导数，对于更新权值的算法，一般选择梯度下降算法。而再此系统中，我使用的TensorFlow框架，只需要设置好参数进行训练就好，烦杂的过程框架都帮你解决好了。同时由于神经网络的初始值的选择在一定程度上影响最终结果和训练的效率，因此我使用遗传算法给初始值进行优化。在这个系统中，我设计的隐藏层有9个神经元，因此结构11-9-3，学习率0.1，训练次数1000次。因此是[11,9]的矩阵，b1是[1,9]的矩阵，是[9,3]的矩阵，b2是[1,3]的矩阵。一共138个参数，因此遗传算法中的个体大小为138，同时设置种群大小为20，遗传代数为20，交叉概率0.75，变异概率0.05。

使用遗传算法优化的BP的伪代码如下：

-------------------------------------------------------------------

算法：GA-BP

-------------------------------------------------------------------

输入：带标签的训练数据集，测试数据集

-------------------------------------------------------------------

输出：训练好的模型,预测的数据

-------------------------------------------------------------------

01：初始化神经网络，学习率，选择权值更新算法

02：定义种群大小m，交叉概率cp,变异概率mp,遗传代数n

03：根据神经网络设计适应度函数

04：for i=0 to m

05: 初始化种群中的个体，将神经网络的权值编码成个体的基因

06：endfor

07：for i=0 to n

08： for j=0 to m

09： 计算每个个体的适应度

10： endfor

11： for j=0 to m

12： 参考适应度根据某种选择选择算法选择交叉对象

13： 根据交叉概率对选择的对象进行交叉生成新个体

14： 对每个新个体依据变异概率进行变异操作

15： endfor

16： if 上一代适应度最高个体 > 当代适应度最差

17： 用上一代适应度最高替换当代最差的个体

18： endif

19：endfor

20：将迭代完的种群中的最优个体的基因性解码得到每个权值的值

21：将权值代入到神经网络模型中

22：for t=0 to 训练次数

23： 代入训练数据进行训练

24：endfor

25：保存训练好的模型

26：将测试集数据放在训练好的模型上进行正向得到预测的值

-------------------------------------------------------------------

对于长短期记忆（LSTM）。我们同样需要设置网络的神经容量，即网络的层数和每层的“神经元”（Block）。以及损失函数，权值更新函数，除此之外还有Block的结构和序列长度。在使用Tensorflow时，每种神经网络的训练基本由其自动完成，因此主要就是对神经网络的结构以及相关的参数进行设计。

对于模型的训练过程，由于训练的过程比较繁琐，在此就没有详细论述，在BP网络训练的初期，神经容量比较小，训练的模型还处于欠拟合，随着神经容量的增大，模型的拟合的效果慢慢变好，在此期间我一共尝试了10种网络结构，从3层到6层，在6层左右时， 部分数据存在过拟合的现象，对比综合性能，最后，选择了三层的网络结构。

同时，将三种方面分别使用三个类进行设计，每个类完成一种方法的预测过程，采用面对对象的方法，对于每个模型的训练，即找到最佳的参数的设置。我们在完成方法对象的设计之前就已经决定好了。因此在最后阶段的集成中，直接调用三个类的对应方法，对于其他的数据集，由于每个测量桩的数据都是类似，因此，训练过程采用的模型的一般参数就使用我们训练桩的参数，虽然存在一定的误差，但应该在可以接受的范围内。

**5.3.3 模型验证功能实现**

对于验证，我们将之前分出来的用于测试的数据放在训练好的模型上运行，根据实际值和测试值的差别，看每个模型的拟合效果，在这里我们使用的是第四章介绍的模型评估方式，为了方便编写，以及简化实验的步骤，在这里我直接使用sklearn库中评估模型。在使用该库中的评估模型时，对于每一个方法，其参数形式为model(y, )，调用该方法直接得出每个模型对应的最终结果，为了更加好的显示我们的结果，我使用了pandas库进行建立回归指标的数据框。将测试集代入之前训练的模型中得到预测值，再将测试集中的对应的实际值y和测试集中的预测值代入上述的评估模型中。以下是遗传算法优化的BP神经网络的模型评估结果，由于数据集没有使用归一化，导致MAE和MSE的值超出正常范围。

表5-1 评估模型



**5.3.4 模型预测功能实现**

对于模型的测试功能，我们先得到我们目前的数据，即11个输入，代入到我们训练好的模型中，得到每个模型的预测值，由于几个模型在有些方面做的差不多，因此我进行的一定的权值进行平均，然后显示在屏幕上，理论上我们得到目前数据的方法是通过传感器进行收集的，由于显示情况中，没法得到实时数据，在此我只能将以前的值作为现实值进行模型。

**5.3.5 系统界面实现**

由于本系统是用Python进行编程的，虽然现在对于在Java中嵌入Python是可行的，但由于在二者之间进行转化比较麻烦，加之Python自己的UI也不错，故在此我使用的是Python的UI，使用的是tkinter库进行UI编程的。整个UI界面使用的2个Frame（left and right）,在left中有4个Button（选择文件、训练、测试、预测），在right中1个Label进行现实图片（训练的成本曲线、验证的对比曲线）,1个Frame用于现实预测的三个值，5，10，15分钟的值。

**5.4 系统展示**

**5.4.1 系统界面展示**



图5-2 界面设计图

**5.4.2 选择文件**

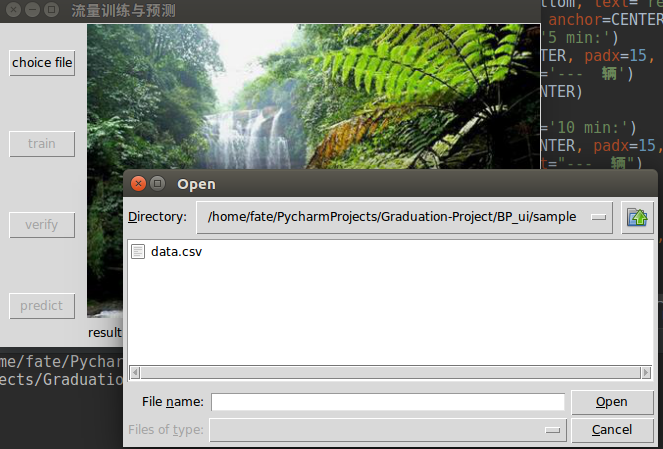


图5-3 选择文件的界面

**5.4.3 模型训练**

系统一个模型的训练误差，训练次数为10000次。

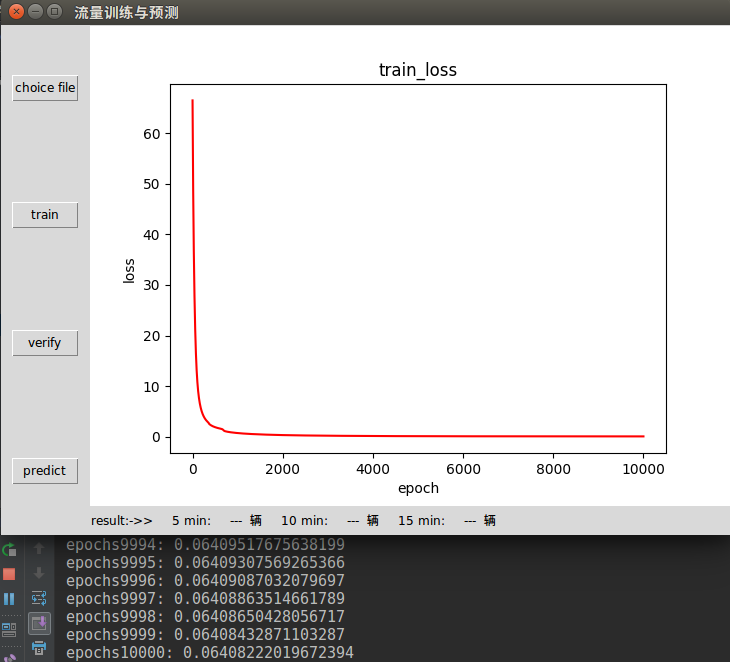


图5-4 训练之后的界面

**5.4.4 模型验证**

将测试集中的数据，导入训练好的模型中，随机选出一天，测试值和正常值的数据对比，X坐标：0-300，是一天的数据，每5分钟一个，一共288个。Y坐标：是四个车道的总的车流量单位是辆。

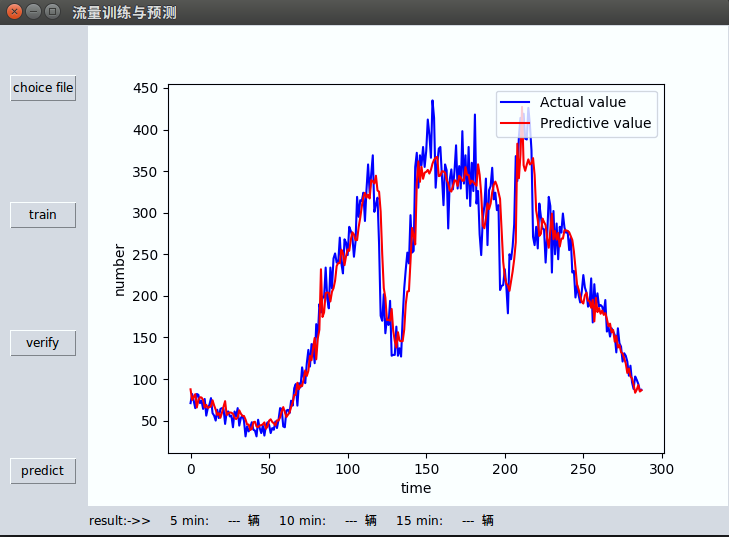


图5-5 模型验证之后的界面

**5.4.5 进行预测**

点击Predict，我们自动提取预测下一个数据所需要的相关信息，导入到训练好的模型中，得到未来5-15分钟的车流量。



图5-6 预测之后的界面

**5.5 实验结果-数据**

由于本文使用了三种模型，在此，对每一个模型，采用基本相同的模型参数，再使用单源和多源数据进行训练，观看每个模型在这两种数据即上的表现。

BP神经网络在两种数据上的表现：

表5-2：单源数据上BP的评估



表5-3：多源数据上BP的评估



SVR在两种数据上的表现：

表5-4：单源数据上SVR的评估



表5-5：多源数据上SVR的评估



LSTM在两种数据上的表现：

表5-6：单源数据上LSTM的评估



表5-7：多源数据上LSTM的评估



**5.6 实验结果-模型**

对于模型的选择上，采用多种模型相结合的方法，理论上可能使得将每个模型的优点结合起来，但实际上，我们需要控制好每个模型的贡献比，以及模型结合的方法，不当的结合方法对最终模型的效果可能起到事倍功半的效果。在这里我们采用随机结合的方式，即使用3种模型预测的值，然后根据一个随机数取三个模型中的一种。在这里就直接使用多源数据。

表5-8：综合模型在多源数据上的评估



**5.7 实验总结**

由上可知，神经网络在多源数据集上的表现比单源的稍微优秀（比较绝对误差），而支持向量回归则相反，一种可能是神经网络的学习能力比支持向量回归要好，另一种原因可能是对不同样本，模型的参数需要进行一定的修改。同时，两类数据上的表现差异不明显，可能由于天气数据不够精确和测量间距不匹配，其次就是，每种模型的最优参数没有训练出来。

**5.8 小结**

本章主要介绍了基于多源数据的高速公路流量预测系统的实现过程，并对系统界面进行了展示。由于自己也是初次接触机器学习方面的东西，对于大量的参数调节工作，自己有时也是一面懵逼，对于BP神经网络的调参，从一开始只知道的对训练次数的调节。再到学习率的动态调节，再到更新权值的方法的选择（常用的是梯度下降），对于模型的欠拟合、过拟合问题的处理等大量工作自己也不是很清楚。接着就是调节网络的结构，从一开始只知道增加隐藏层的神经元、隐藏层的层数等，随着越来越多神经容量，增加的模型出现各种问题，接着就参看别的调节参数的方法，优化BP神经网络的方法等。其实在做毕业设计之前，自己只对BP神经网络有些了解，对于其他的，比如支持向量回归、长短期记忆网络等的了解非常少。因此在实现过程中存在大量的问题，有些能改正，有些则超出自己的能力范围。

**第6章 总结与展望**

从开始接手这个毕业设计，到现在基本实现了一个系统，真的应验了那么一句话，有很多东西你单单去学习是学不到东西的，在项目中进行学习更加有效。一开始接收到的选题是一个3D建模软件的二次开发，由于自己大三暑假期间接触了一些吴恩达老师的机器学习的课程，因此选择了这个选题，从一开始打算只使用BP神经网络进行训练，到调节参数无赖选择其他的方法，什么SVR、LSTM、GBR(梯度增强回归模型)、BR（贝叶斯岭回归模型）、ETC（弹性网络回归模型）等，后面的只是简单的测试了一下，最终只选择了BP、SVR、LSTM。同时使用遗传算法优化BP神经网络，也开始进一步的了解了机器学习方向的东西，这个毕业设计让我学会了很多东西，同时由于不熟练的缘故，也存在大量的不足。希望老师们可以理解。

随着科技的发展，人们追求的生活的方式也有了一定的改变，正如我们以前对支付宝的等在线支付的态度，从一开始的钱包，到现在的只需要拿一个手机就能解决所有的事，这个改变突显了科技的力量，而这个生活态度有反过来促进人们对科技的依赖，随着人工智能的发展。科技一点一点的改变传统的生活方式，与此同时，随着人们对太阳系的探索，人们对地球，对环保又有了新的认识。而如何对地球的保护也吸引了各国的重视，可持续发展成了人们的保护地球的共识。即使用恰当的资源满足自己的需求，提高资源的利用率，不浪费，同时在使用资源的同时，又不能过度。因此这又变成了另一个问题，就是最优化问题，然而不是所有的人，都有能力和时间去完成一个又一个最优化问题，由于计算机强大的计算能力，人工智能慢慢开始“智能”起来，这个问题的解决方法有了理论上的基础。即人们借助人工智能去完成各种“最优化”问题。

结合我们研究的交通，每年都因此堵车，事故等原因造成了大量的人力、物力的浪费，同时堵车造成的环境污染、资源浪费的问题也不容忽视。虽然目前的有许多研究电动汽车的公司，但价格完全民众化还存在一段距离。因此对“公路资源”进行“最大化”研究是一个值得深究的问题。借助科技的发展，使用更多，更发达的数据采集器，通过人工智能进行最优化处理，是历史发展的必然趋势。

智能城市是这个应用的一个雏形，虽然目前的显示还有待优化，当一旦其成熟，对环境、资源的保护是有很大的作用的。单单一个交通流量的预测，对交通拥堵问题就有一定的缓解，同时对无人驾驶进行最优化路径的选择也有一定的帮助。

因此，我相信对于交通这领域，会有更多人工智能的因素加入其中。

**致 谢**

在做本次毕业设计的过程中，我结合之间看的吴恩达老师的机器学习方面的知识结合起来，从实践中再将理论知识理解一遍，有些之前没有理解的东西，也开始慢慢熟悉起来了，并在实践过程中逐渐理解并将其运用到本次回归系统的实现中，同时也为未来学习上遇到类似问题提供多种选择。

总以为自己很难完成这个任务，完成以后，首先要感谢河海大学谢在鹏老师的悉心指导。每过一段时间就询问我的毕设进展情况，并针对毕设中存在的问题，给出解决问题的建议，以及对忘记补充的内容进行补充。在论文写作期间，对我进行督促，而且还提供各种编写文档的工具，很感谢老师的热心帮助，同时也要感谢好基友景宗飞，在暑假期间推荐我看吴恩达的机器学习相关课程。

然后还要感谢宿舍的兄弟，他们在我做毕业设计的过程中提供了安静、和谐的宿舍环境，以及好朋友对我毕业设计的建议和看法，这些都对我完成毕业设计都有着很大的帮助。

最后，还要感谢东南大学的杨明老师、以及JSSEC实验室一里面的前辈。

参考文献

[1] 公安部统计机动车新用户数据.

[2] 腾讯互联网+ 什么是智慧城市 http://plus.tencent.com/detailnews/824.

[3] 王军,许宏科,蔡晓峰,孙磊.基于BP神经网络的高速公路动态交通流预测[J].公路交通技术,2007(01):150-152+163.

[4] Papageorgiou Markos,Blosseville Jean- Marc,Hadj- Salem Habib.Modeling and real- time control of traffic flow on the southern part of boulevard peripherique in Paris[ J]. Transportation research,1990,24A(5):345- 359.

[5] 韩惠婷.高速公路短时交通流量组合预测[J].江苏科技信息,2013(04):71-74.

[6] 王宏杰,林良明,徐大淦,颜国正.基于改进BP网交通流动态时序预测算法的研究[J].交通与计算机,2001(03):11-14.

[7] 陈玛瑙. 面向城市路网的短时交通参数预测方法研究[D].吉林大学,2017.

[8] 段晋茂,潘玉民.粒子群优化的神经网络预测交通流量[J].黑龙江科技信息,2013(19):83-85.

[9] 李松，刘力军，解永乐.遗传算法优化BP神经网络的短时交通流混沌预测[J].控制与决策，2011,26(10):1581-1585．

[10] 卢建中,程浩.改进GA优化BP神经网络的短时交通流预测[J].合肥工业大学学报(自然科学版),2015,38(01):127-131.

[11] 刘彩云. 基于Spark的改进SA-SVR短时交通预测研究[D].东华理工大学,2017.

[12] 葛志鹏. 基于多源数据的高速公路短时交通状态预测方法研究[D].东南大学,2016.

[13] 王秋莉,李军.基于核学习方法的短时交通流量预测[J/OL].计算机应用研究,2019(03):1-2[2018-05-10].

[14] 刘瑞. 基于人工蜂群算法和神经网络的交通流量预测研究[D].青岛理工大学,2018.

[15] 戴坤成,邱志敏.ICWLS-SVM在城市道路短时交通流量预测建模中的应用[J].兰州工业学院学报,2018,25(01):78-83.

[16] 张利,李星毅,施化吉.基于ARIMA模型的短时交通流量预测算法研究[J].郑州轻工业学院学报(自然科学版),2008(04):89-92.

[17] 刘驰,刘伟铭.基于小波变换的交通流预测及模型优化[J].科学技术与工程,2008(21):5858-5862.

[18] 刘祥.基于贝叶斯网络的多方法短时交通流量联合预测[J].机电技术,2010,33(04):33-34+45.

[19] M. S. Ahmed and A. R. Cook, “Analysis of freeway traffic time-series data by using Box–Jenkins techniques,” Transp. Res. Rec., no. 722, pp. 1–9, 1979.

[20] Chenyun Yu,Ka Chi Lam. Applying multiple kernel learning and support vector machine for solving the multicriteria and nonlinearity problems of traffic flow prediction[J]. Journal of Advanced Transportation,2014,48(3).

[21] Chun Mei Zhu,Chang Peng Yan,Xiao Li Xu,Guo Xin Wu. Research on the Application of the Prediction of the Expressway Traffic Flow Based on the Neural Network with Genetic Algorithm[J]. Advanced Materials Research,2011,1165(189).

[22] Yiming Li. Traffic Flow Prediction With Big Data: A Learning Approach Based on SIS-Complex Networks[A]. IEEE Beijing Section、Global Union Academy of Science and Technology、Chongqing Global Union Academy of Science and Technology.Proceedings of 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC 2017)[C].IEEE Beijing Section、Global Union Academy of Science and Technology、Chongqing Global Union Academy of Science and Technology:,2017:5.

[23] 刘光聪Tensorflow架构与设计 变量初始化 https://www.jianshu.com/p/bebcdfb74fb1

[24] 戴红良.基于BP神经网络的高速公路节假日短时交通流量预测研究2015 (46) :236-238

[25] 周志华《机器学习》6.1章节 河海大学孟老师《人工神经网络》课程PPT《ANN·SVM》

[26] 简书石晓文的学习日记SVM系列十三讲--支持向量回归机SVR https://www.jianshu.com/u/c5df9e229a67

[27] 邹恒明《算法之道第二版》第15.11章遗传算法.

[28] 河海大学孟朝晖《人工神经网络》课程PPT《ANN part1 print》

[29] 周志华《机器学习》5.1章节神经元模型

[30] 吴恩达网易云课堂上的公开课《人工神经网络》微课程内容.

[31] iterate7博客LSTM和循环网络RNN学习简记

<https://blog.csdn.net/iterate7/article/details/76301852>.

[32] colah一篇《Understanding LSTM Networks》的博客.

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[33] 河海大学孟朝晖《人工神经网络》课程PPT《ANN part2 print》