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 博士 |  |
| **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

交通流量预测是测量交通状态的一种方式， 实时的交通状态对交通控制和智能交通有重要意义。交通流量预测一般分为三种：短期、中期、长期。由于交通流量具有波动性，出于预测的精确度，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，并将其应用于预测未来数据上。大量长期有效的数据是实现这一目标的基础。传统的交通流量预测主要是根据历史的交通流量进行分析，而忽视交通流量在一定程度上的相关性，导致其输入单一。为了提高交通流量预测的准确性，用多源数据进行短时交通流量预测，增加各种影响因素，从而提高准确率。多源数据，例如，环境因素（降雨量，温度，光照），收费站进出口数。

由于交通流量的非线性，同时神经网络在处理非线性问题上有着先天的优势。【Hornik et al.,1989】证明，只需一个包含足够多神经元的隐层，BP网络就能以任意精度逼近任意复杂度的连续函数。此外SVR，集成学习等方法在非线性处理上也有一定效果。因此在算法的选择上，结合多种优质算法进行预测。在研究对象的选择上，由于高速公路的路网结构相对其他道路简单，相关制约因子容易理解。故选择高速公路作为实践对象。

因此，本文将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。并制作了一个简单的UI进行相关操作。

短期流量预测，对智能交通系统的发展有重要意义，而其预测的精确度更是衡量智能交通系统的指标。

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

**ABSTRACT**

Traffic flow forecasting is a way of measuring traffic conditions. Real-time traffic conditions are important for traffic control and intelligent traffic. Traffic flow prediction is generally divided into three types: short-term, medium-term, and long-term. Due to the volatility of traffic flow, it is more meaningful to study short-term traffic flow for the accuracy of the prediction.

Short-term traffic flow forecasting can be viewed as a non-linear regression problem. Historical rules can be traced out or learned as various parameters for learning a fitting function, and used to predict future data. A large amount of long-term effective data is the basis for achieving this goal. The traditional traffic flow forecasting is mainly based on the analysis of historical traffic flow, while ignoring traffic flow to a certain degree of correlation, resulting in a single input. In order to improve the accuracy of traffic flow forecasting, short-term traffic flow forecasting using multi-source data, Improves accuracy by various influencing factors. Multi-source data, for example, environmental factors (rainfall, temperature, light), import and export number of toll stations.

Due to the nonlinearity of traffic flow, neural networks have inherent advantages in dealing with non-linear problems. [Hornik et al., 1989] demonstrated that a BP network can approximate a continuous function of arbitrary complexity with arbitrary precision with only one hidden layer containing enough neurons. In addition, methods such as SVR and ensemble learning also have certain effects on nonlinear processing. Therefore, combining a variety of high-quality algorithms for prediction, the choice of research objects, because the road network results of the highway is relatively simple compared to other roads, and the relevant constraints are easy to understand. Therefore, the highway is chosen as a practical object.

Therefore, this paper will use multi-source data for short-term traffic forecasting and implement a short-term traffic forecasting system based on a combination of multiple regression algorithms based on BP neural networks.

Short-term traffic forecasting is of great significance to the development of intelligent transportation systems, and the accuracy of its predictions is a measure of the performance of intelligent transportation systems (ITS).

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

**目录**

[摘 要 I](#_Toc513463639)

[ABSTRACT II](#_Toc513463640)

[目录 IV](#_Toc513463641)

[第1章 绪论 6](#_Toc513463642)

[**1.1研究背景** 6](#_Toc513463643)

[**1.2国内外研究现状** 8](#_Toc513463644)

[**1.3本文工作** 9](#_Toc513463645)

[第2章 TensorFlow框架、提取预测的样本形式 10](#_Toc513463646)

[**2.1 TensorFlow简介** 10](#_Toc513463647)

[**2.2 交通状态** 13](#_Toc513463648)

[**2.3 提取样本** 14](#_Toc513463649)

[**2.3.1 数据准备** 14](#_Toc513463650)

[**2.3.2 工具介绍** 14](#_Toc513463651)

[**2.3.3 提取步骤** 14](#_Toc513463652)

[**2.4 小结** 23](#_Toc513463653)

[第3章 短时流量预测的相关算法 24](#_Toc513463654)

[**3.1 支持向量回归** 24](#_Toc513463655)

[**3.1 遗传算法** 27](#_Toc513463656)

[**3.3 BP神经网络** 29](#_Toc513463657)

[**3.4 长短期记忆网络** 32](#_Toc513463658)

[**3.5 小结** 38](#_Toc513463659)

[第4章 基于多源数据的高速公路流量预测系统的设计 39](#_Toc513463660)

[**4.1 需求分析** 39](#_Toc513463661)

[**4.2 总体设计** 39](#_Toc513463662)

[**4.2.1 系统结构设计** 39](#_Toc513463663)

[**4.2.2 系统流程设计** 39](#_Toc513463664)

[**4.3详细设计** 41](#_Toc513463665)

[**4.3.1 处理输入数据** 41](#_Toc513463666)

[**4.3.2 定义模型参数** 42](#_Toc513463667)

[**4.3.3 确定模型评估方式** 42](#_Toc513463668)

[**4.3.4 界面设计** 43](#_Toc513463669)

[**4.5 小结** 44](#_Toc513463670)

[**5.1 系统开发环境** 45](#_Toc513463671)

[**5.2 系统实现** 45](#_Toc513463672)

[**5.2.1 输入数据功能实现** 45](#_Toc513463673)

[**5.2.2 模型训练功能实现** 45](#_Toc513463674)

[**5.2.3 模型验证功能实现** 46](#_Toc513463675)

[**5.2.4 模型预测功能实现** 47](#_Toc513463676)

[**5.2.5 系统界面实现** 47](#_Toc513463677)

[**5.3 系统展示** 47](#_Toc513463678)

[**5.3.1 系统界面展示** 47](#_Toc513463679)

[**5.3.2 选择文件** 48](#_Toc513463680)

[**5.3.3 模型训练** 48](#_Toc513463681)

[**5.3.4 模型验证** 48](#_Toc513463682)

[**5.3.5 进行预测** 48](#_Toc513463683)

[**5.4 小结** 48](#_Toc513463684)

[第6章 总结与展望 48](#_Toc513463685)

[致 谢 50](#_Toc513463686)

[参考文献 51](#_Toc513463687)

**第1章 绪论**

**1.1研究背景**

近年来，随着我国物质生活水平的提高，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，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3.10亿辆，其中汽车2.17亿辆；机动车驾驶人达3.85亿人，其中汽车驾驶人3.42亿人），驾车外出的情况也逐年爬升。根据公安局的统计图显示，这种情况还在不断加剧。

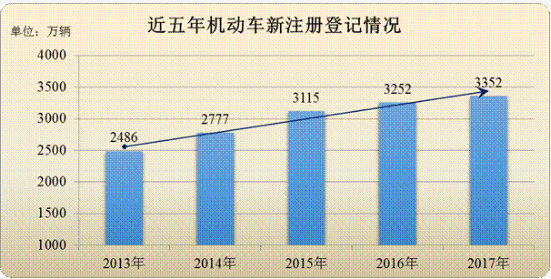


图1 机动车新用户

因此交通公路承受的压力也在不断加大。虽然我国公路也在不断发展，从建国初的几万公里到2017的400万公里。但是严重的堵车现象还是在局部地区出现，看似这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。于此同时，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收到人们的关注。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。不满足社会人的东西终究会被遗忘在历史的长河中，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。这导致在某些路段上，交叉路口的顺畅程度极不对称。而人们渴求的是一个能动态调整的交通系统。

智能交通系统（ITS）随之而出，根据2018年的一篇来自腾讯互联网+的文章，智能交通系统的概念如下：先进的科学技术（信息技术、计算机技术、数据通信技术、传感器技术、电子控制技术、自动控制理论、运筹学、人工智能等）有效地综合运用于交通运输、服务控制和车辆制造，加强车辆、道路、使用者三者之间的联系，从而形成一种保障安全、提高效率、改善环境、节约能源的综合运输系统。

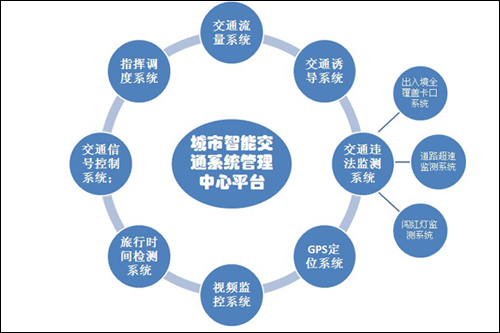


图2 智能交通系统概率图

其作用通过人、车、路的和谐、密切配合提高交通运输效率，缓解交通阻塞，提高路网通过能力，减少交通事故，降低能源消耗，减轻环境污染。

通俗地讲，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），在北京、上海、广州等大城市已经建设了先进的智能交通系统；其中，北京建立了道路交通控制、公共交通指挥与调度、高速公路管理和紧急事件管理的4大ITS系统；广州建立了交通信息共用主平台、物流信息平台和静态交通管理系统的3大ITS系统。

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。交通流量系统的主要作用就是对交通状态的把控。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，交通流的特征参数：交通速度、交通流量、路段占有率、交通流密度。而一般情况下交通速度在既定的道路上是一定的，当交通流量达到一定时，其速度才会受到影响，同理路段占有率，交通流密度都随着交通流量的变化而变化。因此大部分研究交通流的都是处理交通流量。已经交通流量对交通状态进行评估和预测。

交通流量，即根据各种数据，进行交通流量预测。根据收集到的有关交通的相关因子，通过各种手段，进行流量预测，得到未来的交通流，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流量的预测，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，主要有以下几种：人工输入、GPS车载导航仪器、GPS导航手机、车辆通行电子信息卡、CCTV摄像机、红外雷达检测器、线圈检测器、光学检测仪等等。随着大数据和硬件技术的发展，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。交通流量预测，用历史数据和实时数据来预测未来时段的交通流量。由于其研究为驱动型研究，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。

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，相关工作越来也精确。而交通流量预测可归结为回归问题，常见的回归问题解决方案有，基于统计理论：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；基于非线性预测理论：小波理论、基于突变理论、基于混沌、分形理论；基于人工智能理论：人工神经网络、SVM；基于组合混合模型：小波理论混合、神经网络混合（神经网络、模型系统、时间序列、遗传算法）；基于微观交通仿真的模型。

王军、许宏科等在Payne的MACK模型【1】上基于BP神经网络进行的高速公路动态交通流预测【2】，其中介绍了神经网络的输入层、隐藏层、输出层的个体数目。在对比RBF模型后，得出其可靠性更高。

韩惠婷基于自适应指数平滑和RBF神经网络的组合预测模型模型【3】，王宏杰、林良明等提出调整学习率和在修改权值中增加动量的改进的BP神经网络进行流量预测【4】，陈玛瑙提出的基于灰色关联度分析的Kalman滤波算法【5】，卢建中、程浩提出的使用改进的GA算法优化BP神经网络的短时交通模型【6】。

东南大学交通运输规划与管理葛志鹏采用多源数据对交通状态的预测【7】，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。

**1.3本文工作**

本文的内容将从以下六个方面展开：

第1章 概述。主要介绍了短时交通流量的研究背景，国内外研究现状及相关技术的发展和应用情况，并对论文的组织结构进行了介绍。

第2章 提取短时流量预测的样本以及Tensorflow框架。

第3章 短时流量预测的相关算法。本章详细介绍了本文需要用到的SVR，LSTM、GA算法优化的BP神经网络进行了学习，并用数据进行了实验分析，为系统开发提供理论支持。

第4章 短时交通流量预测的设计。本章介绍系统的需求分析、总体设计、数据库设计、功能模块、界面设计等，为系统实现提供模版。

第5章 短时交通流量预测系统的实现。本章简单介绍了系统的开发环境，对系统的实现进行了详细的说明，并展示了系统的操作流程。

第6章 本章对本论文完成的工作进行了总结，说明了论文所完成的研究，提出了不足之处以便以后的改进。

**第2章 TensorFlow框架、提取预测的样本形式**

**2.1 TensorFlow简介**

随着围棋大赛中人机大战的一波又一高潮，TensorFlow慢慢浮出水面，它是由Google公司Google Brain团队开发的一个开源软件库，用于各种感知和语言理解任务的机器学习。最初用于Google研究和生产，于2015年11月9日在Apache2.0开源许可证下发布。

从2010年开始，Google Brain建立DistBelief作为他们的第一代专有的机器学习系统，50多个团队在Google和其他Alphabet公司在商业铲平部署了DistBelief的深度学习，神经网络。包括Google搜索、Google相册等，由于随着原始代码库的不断发展，以及机器学习的空前大热，Google指派计算机科学家，如Geoffrey Hinton和Jeff Dean，简化和重构DistBelief的代码库，使其变成一个更快、更健壮的应用级别代码库，如是TensorFlow横跨出世。2009年，Hinton领导的研究小组大大减少使用DistBelief的神经网络错误的错误数量，通过Hinton在广义反向传播的科学突破，使得Google语音识别软件中的错误减少至少25%。

作为Google Brain的第二代机器学习系统，1.0.0版本发布于2017年2月11日，虽然参考实现运行在单台设备，TensorFlow可以运行在多个CPU和GPU（和可选的CUDA扩展和图形处理器通用计算的SYCL扩展）。TensorFlow可用于64位Linux、macOS和Windows，以及移动计算平台，包括Android和iOS。

TensorFlow是一个采用数据流**图**(data flow graphs)，用于数值计算的开源软件库，**节点**(Node)在图中表示数学操作，图中的线(edges)则表示在节点间相互联系的多维数据组，即**张量**(tensor)。Graph表示了数据流图中的操作流水，Tensor表示其中流动的数据。TensorFlow程序通常被组织成一个构建阶段和一个执行阶段。在构建阶段操作的执行步骤被描述成一个图，在执行阶段，使用会话执行执行图中的操作。使用这种结构方便计算后面神经网络的反向传递。

构建一个简单的图：

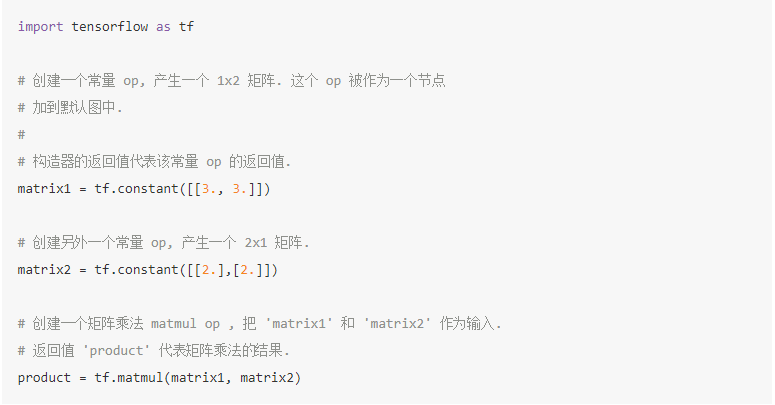


图3 tf中构建计算图的代码

此时的product只是一个Tensor，无法输出相对应的值，只是构建了一个计算图。

执行这个图：



图4 tf中执行计算图代码

想了解更多TensorFlow的知识推荐去其官网，作为一个小白，其实自己也是摸着石头过河。下面引用简书（[\_RayCloud\_](https://www.jianshu.com/u/49d1f3b7049e)）的一种理解：

构建：

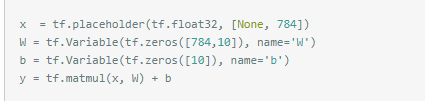


图5 tf中构建计算图

执行：

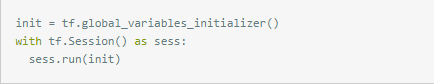


图6 tf中执行计算图

计算图：

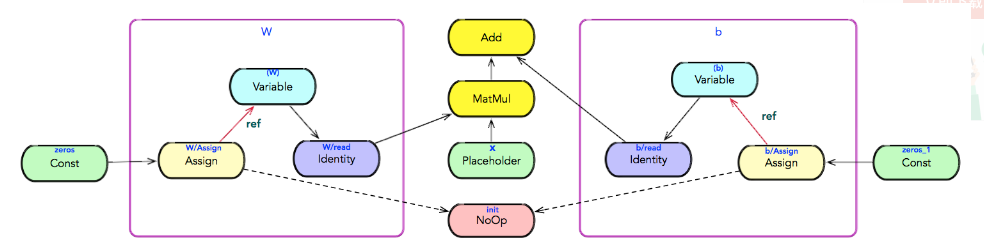


图7 构建的计算图模型

**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向表示交通状态。

交通流是指汽车在道路上连续行驶形成的车流。广义上还包括其他车辆的车流和人流。根据交通流的参数类型，可分为宏观参数：流量、流速、速度和交通流密度等，微观参数：车头时距和车头间距等。然而对于交通流的预测，一般对更能表现的宏观数据进行研究。即流量、流速、速度和交通流密度。

流量，单位时间经过交通道路上指定地点或断面的车量数。流速，即指点地点或断面一小时的流量。交通流理论是运用物理和数学的定律来描述交通特性的一们学科，作为交通工程学的基础理论，广泛用于交通规划、交通控制道路与交通工程设施设计等方面。对于智能交通系统的一部分，进行交通流量系统的设计，因此需要借用交通流理论定量地秒速交通流。根据交通流理论，定量描述交通流的三个参数：交通流量、交通流速度、交通流密度。其中：

交通流量=交通流速度X交通流密度

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预测，从侧面描述交通状态。而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 提取样本**

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。

**2.3.1 数据准备**

本文的交通数据来自于交通局，一共14个测量桩，记录着来回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。车距，车速以此来推。记录时长从2017-06至2017-08共三个月。由于基于多源，因此除了流量，还有天气、上下游相关收费站收据。对于天气，其数据来源于测量桩附件的气象局。对于收费站，一样来之于交通局。（最后发现收费站数据时间失效与交通流量不匹配，故舍弃）

**2.3.2 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，且该文件问.xlsx,免得再Java或其他语言中对Excel进行操作，所以选择导入数据库，因此用到MySQL。对于天气收据，由于其来之于互联网上，因此可以通过爬虫进行爬取。因此再Python中使用bs4，selenium（由于该数据是动态加载的）

**2.3.3 提取步骤**

1、提取数据流信息。

安装MySQL,打开MySQL,如下示：

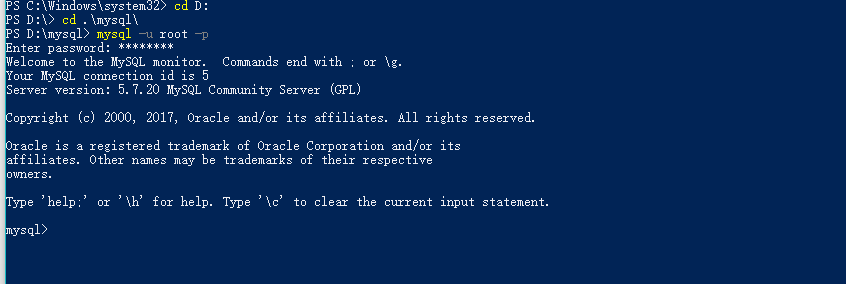


图8 PowerShell中打开MySQL

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：

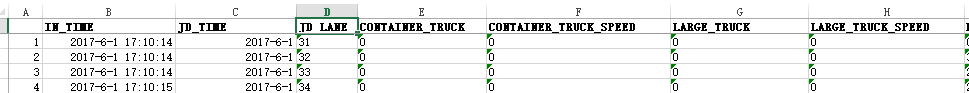


图9 原始的Excel数据（部分）

将Excel中的数据删除没用的列，再另存为.csv格式

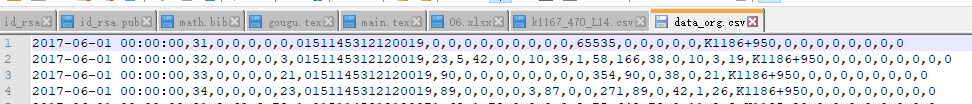


图10 删除几个没有列形成的.csv文件

在数据库中建相应的表：

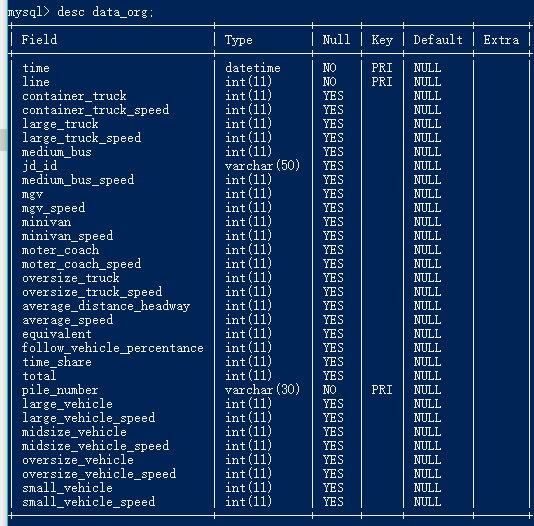


图11 数据库中用于导入的表的列属性

导入外部文件:

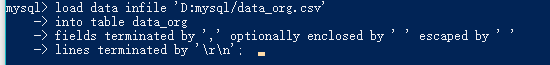


图12 导入代码文件的指令

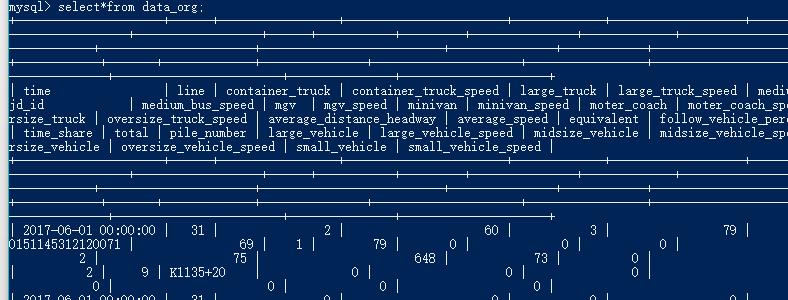


图13 导入之后的结果

再进行简单的SQL语句选择自己需要的数据，导出。

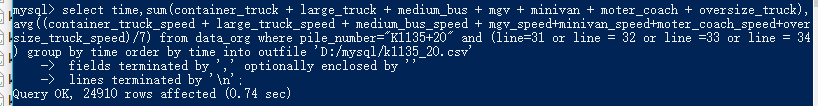


图14 查询车流量和车速

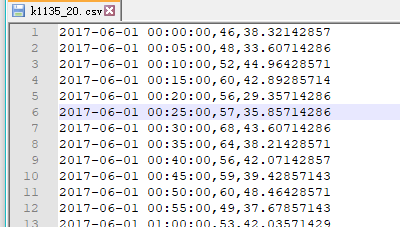


图15 导出的结果

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。而想要历史数据和免费获取，在寻找了很久之后，只有收集历史一天单位的数据（免费情况下）。

用数据里面测量桩编号为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件，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16 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，如下示：



图17 重要内容对应的标签

因此直接使用BeautifulSoup进行搜寻就可以了，如下：

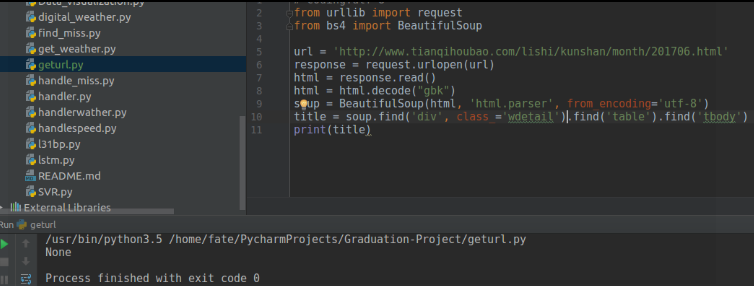


图19 静态提取天气

使用BeautifulSoup查找我们需要的内容，即我们需要的昆山2017年6月份的天气时，出现None,再查找我们主题内容div,就是包含我们需要的数据的更大部分时，显示如下：

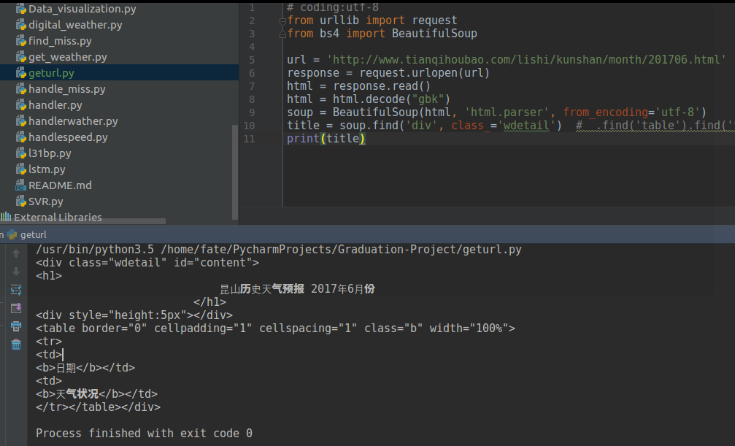


图20 提取div中内容

经过分析得出，这段内容可能时动态加载的，因此使用库selenium，如下：

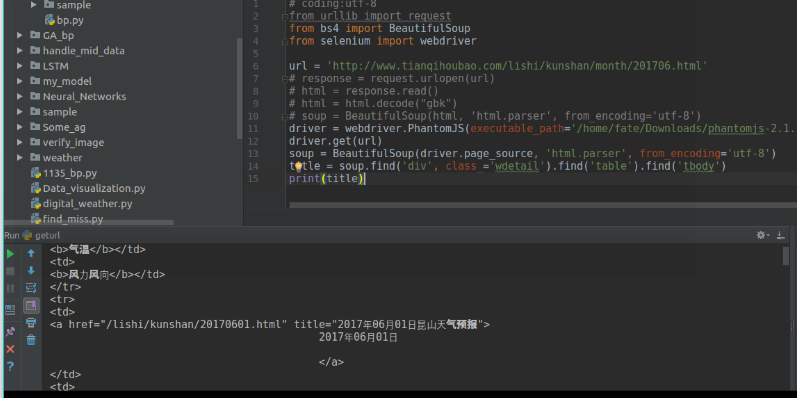


图21 动态提取天气

再把提取好的数据进行相应处理，如下显示：

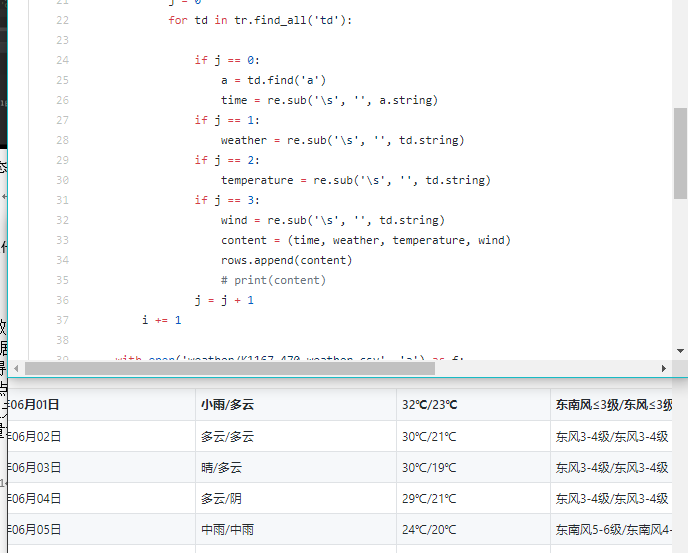


图22 处理代码和结果

2.2.4 合成样本

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【8】，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：

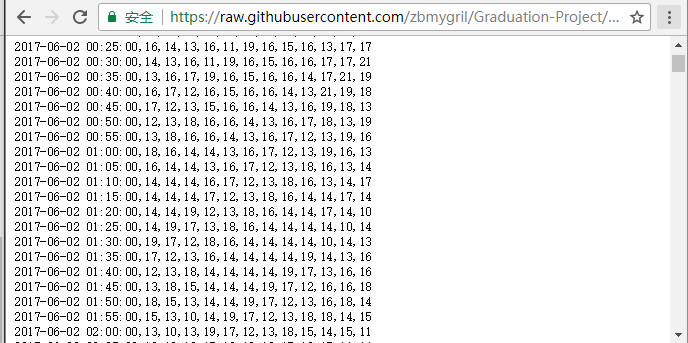


图23 交通流数据



图24 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 。

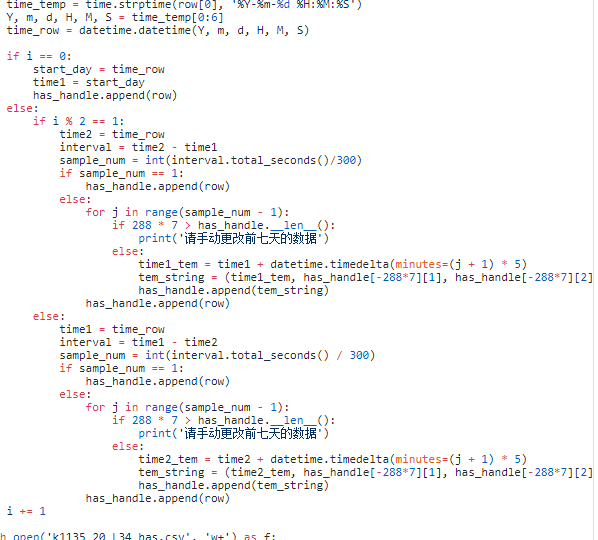


图25 处理丢失样本的代码



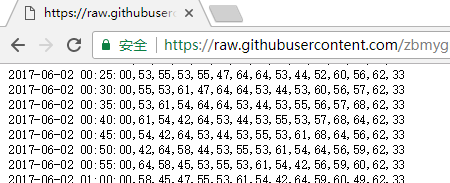
图26 合并交通流量和天气

图27 最终的样本格式

**2.4 小结**

本章简要介绍了TensorFlow的历史和其简单的机制，以及简单的例子。说明测量交通流量的目的是去间接测量交通状态的。详细介绍构建预测样本的过程。为下实现预测模型提供了数据支持。

**第3章 短时流量预测的相关算法**

数据挖掘问题，是通过计算机辅助来挖掘和分析大量数据，从海量的数据中搜索出有趣的模式和重要规律的问题，而机器学习是一个有关对学习过程中的计算方法的研究，以及如何应用基于计算机的学习系统解决实际问题的科学领域。其重要内容时关于在计算机上从数据中产生“模型”的算法，即从数据中学习某种“算法”，在新的数据中运用这个“算法”，在一定的程度上能得到正确的结果，而交通流量预测问题，可以看着数据挖掘问题，即从数据中找到流量的一中规律，在未来用这种规律进行推断。而机器学习则更进一步通过了一种方法，使用存在的“模型”，而不是规律进行指导人们根据目前的数据，得出“结论”。

机器学习，是一门很深厚的学科，它处于计算机科学，工程学以及统计学的交叉点上。要想彻底弄明白需要很长时间，因此本章之间简单介绍一下，机器学习中的支持向量回归（SVR）、遗传算法（GA）、BP神经网络、循环神经网络中的一种长短时间记忆（LSTM）。

**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），在处理分类问题时：

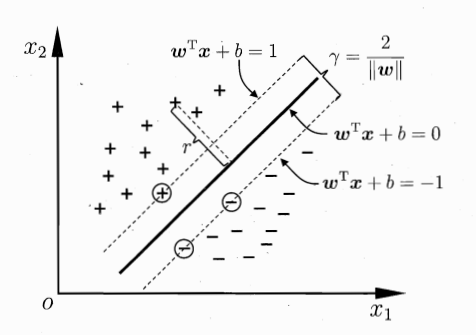


图28 SVM（截选周志华的机器学习）

通过中间的个由两天线组成的隔离代，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，即最大，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着两个参数，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Yi ( WT Xi + b )，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，Yi ( WT Xi + b )1。因此想要得到理想的模型，应该需要满足以下两个条件：

s.t. Yi ( WT Xi + b )1

而支持向量回归（SVR），在处理回归问题时：

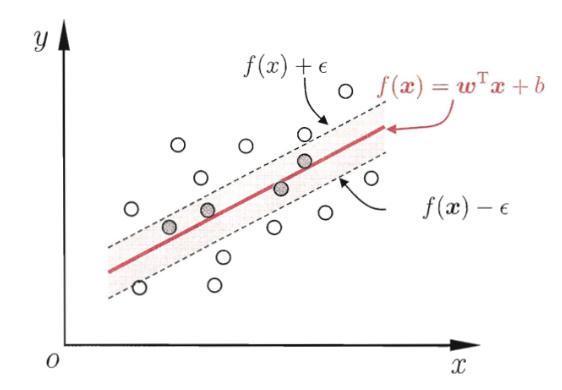


图29 SVR（知乎用户：石晓文的学习日记）

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，因此我们定义一个错误的计算式：

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：

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：

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，通过引用两个变量、：

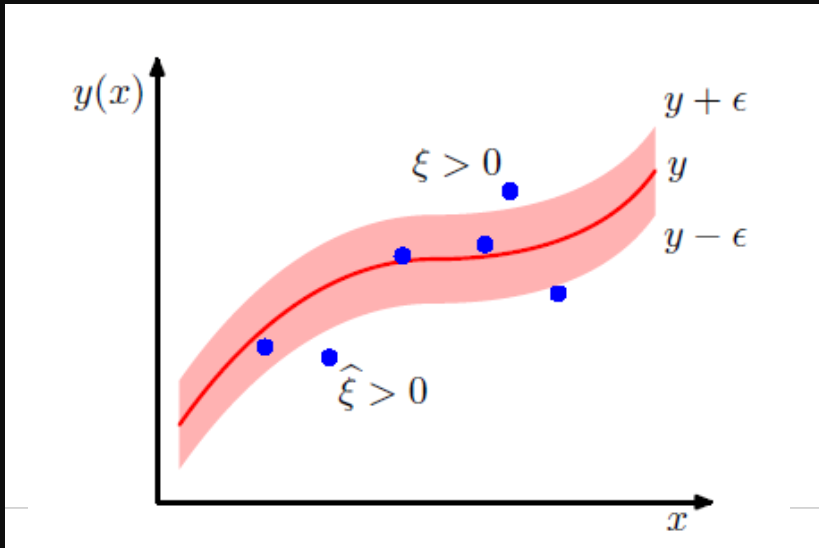


图30 引入松弛变量后的SVR（知乎用户：石晓文的学习日记）

即在原始的间隔带两侧加入松弛带，如是错误表达式如下：

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：

s.t.

,

然后根据拉格朗日乘子，引入对偶问题，根据KKT等进行求解。

**3.1 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）。

遗传算法是一种借鉴生物界的进化和遗传现象的随机化搜索算法，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》著名的结论：优胜劣汰，适者生存。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，以及基因和表现型之间的关系：解码，以及表示解法的个体、种群。

综上所以，在一个遗传算法中，我们需要理解的概率就有：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，再初始化一个种群，根据实际问题设计一个计算适应度的函数，计算该种群中每个个体的适应度，通过选择操作，选择适应度高的个体，再对种群中适应度高的个体进行的繁衍（交叉、变异）。再将父辈中适应度最高的替换新生代中最差的个体（复制）。再查看新生代中的最优个体是否满足当初的要求。若没有达到则重复进行选择、繁衍。

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。一般解的表示有以下几种：二进制编码、互换编码、树形编码。

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：

（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种群，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31 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，即直接将复配的个体，复制到新种群中。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，将上一代两个扩展到多个个体的交配操作。

（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，1943年，McCulloch和Pitts提出第一个神经元数学模型(M-P神经元模型)，并从原理上证明了神经网络能够计算任何算数和逻辑函数，1949年，Hebb提出生物神经元学习的机理，即Hebb学习规则，1958年，Rosenblatt提出感知机网络模型和其学习模型，1960年，Widrow和Hoff提出自适性线性神经元模型和最小均方差学习算法，随着多层神经元的学习变得越来越复杂，人们开始转向其他方向，1986年，Rumelhart等报告了反向传播(BP)算法在神经网络中的应用，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取得的成功，以及Google在神经网络方面的成功，神经网络等开始大火。

生物神经网络结构如下：

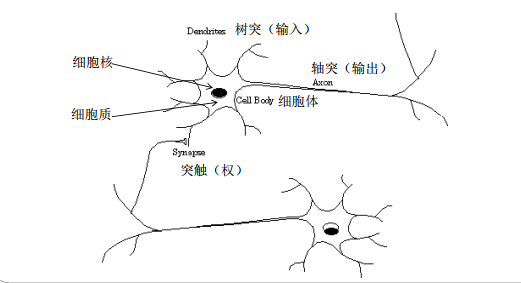


图32 生物神经网络

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如下:

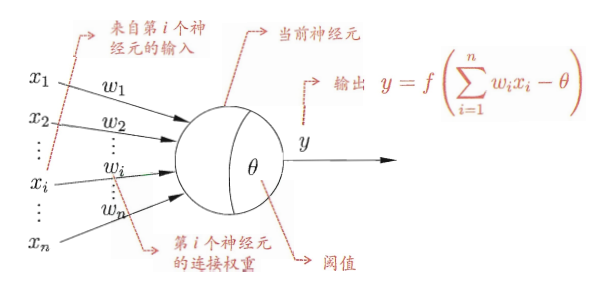


图33 M-P神经模型（截选至周志华老师的机器学习）

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”。

由于BP算法的出现，使得多层网络的学习得到了解决，在一般我们称之为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：

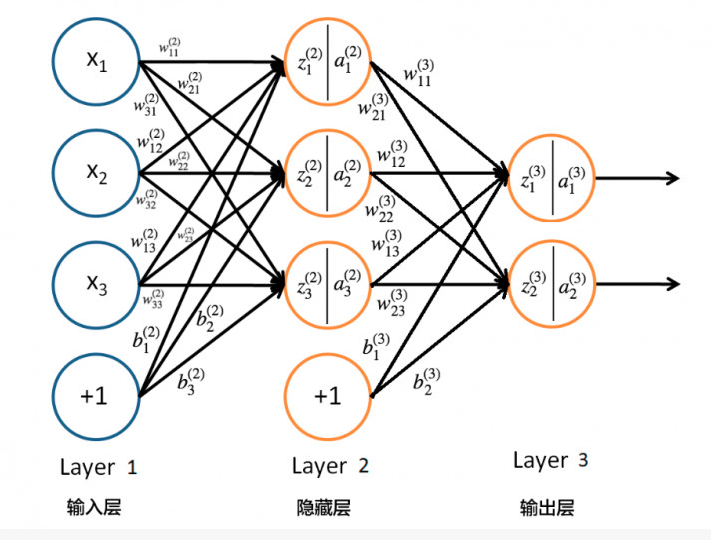


图34 BP神经网络模型（截选自雷锋网Al研究社）

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。

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：

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：

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：

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，传统方面在语音、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。而长短时记忆网络（LSTM）是在循环神经网络的基础上发展起来的一种更新的方法。

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。

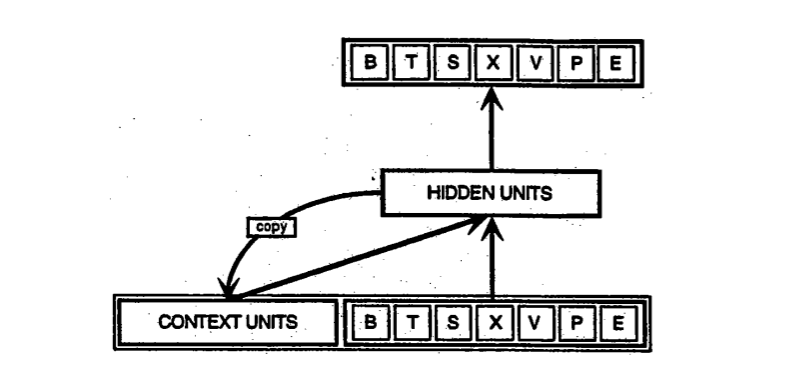


图35 Elman早期RNN的示意图

上图中，BTSXPE 代表当前的输入样例，而 CONTEXT UNIT 则表示前一时刻的输出。

借鉴Christopher Olah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。

Long Short Term Memory 网络通常叫LSTM，是一种特殊的RNN类型，解决RNN梯度消失的问题，LSTM 由Hochreiter & Schmidhuber (1997)提出，并在近期被Alex Graves进行了改良和推广。在很多问题，LSTM 都取得相当巨大的成功，并得到了广泛的使用。下面是其中一个流行的 LSTM 变体，由 Gers & Schmidhuber (2000) 提出的，增加了 “peephole connection”。是说，我们让 门层 也会接受细胞状态的输入。其的示意图：

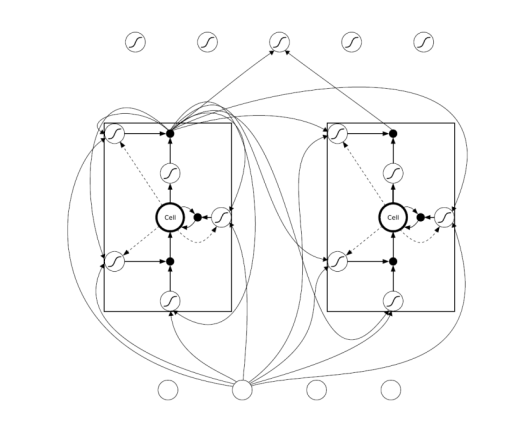


图36 LSTM网络（河海大学孟老师课上PPT）

相比传统的神经网络，其中的神经元被一个黑盒子（Block）抱住了，正是这个Block解决RNN长期依赖的问题。下面我们来看这个Block的结构。

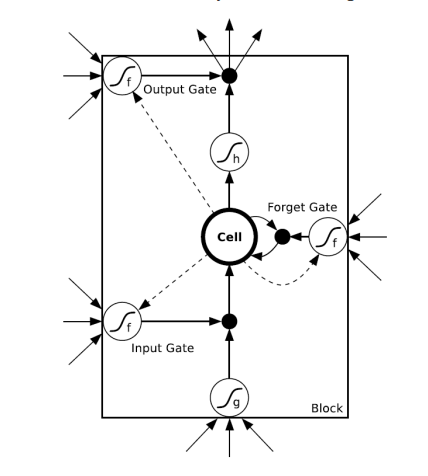


图37 Block单元（河海大学孟老师课上PPT）

每个block包含一个或多个自连接存取单元(Cell)和三个乘法控制器Input Gate、Forget Gate、Output Gate（小黑圈）。三个门是非线性求和单元，用于从块的内部和外部收集激活，并通过相乘点（小黑点）控cell的激活。输入和输出门将神经元的输入和输出相乘，而忘记门乘以单元的先前状态。cell内不应用激活功能。门激活函数'f'通常是logistic sigmoid，所以门激活在0之间（门关闭）和1（门打开）。单元格输入和输出激活函数（'g'和'h'）通常是tanh或logistic sigmoid，尽管如此一些情况'h'是身份函数。虚线表示从Cell到各种门的加权“窥视孔”连接。块内的所有其他连接都未加权（或等同于1.0的固定权重）。从块到网络其余部分的唯一输出从神经元输出和输出门相乘，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。总结一下，LSTM的基本单元Black包含的内容：

3个控制门，input gate输出一个乘法系数控制 cell input 传给 cell 的值,forget gate输出一个乘法系数控制 cell state 的更新程度, output gate输出一个乘法系数控制 cell output 传给 block output 的值。

1组存储器，包含一个或多个cell，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 input gate, forget gate, output gate, cell input,以及输出层或下一个层次。

现在LSTM的基本内容已经了解了，我们来看其是如何工作的，同样在研究了一下DL4J关于LSTM的介绍，下面先从最基本的LSTM开始，其中没有Cell与门之间的连接权，block中的运行如下：

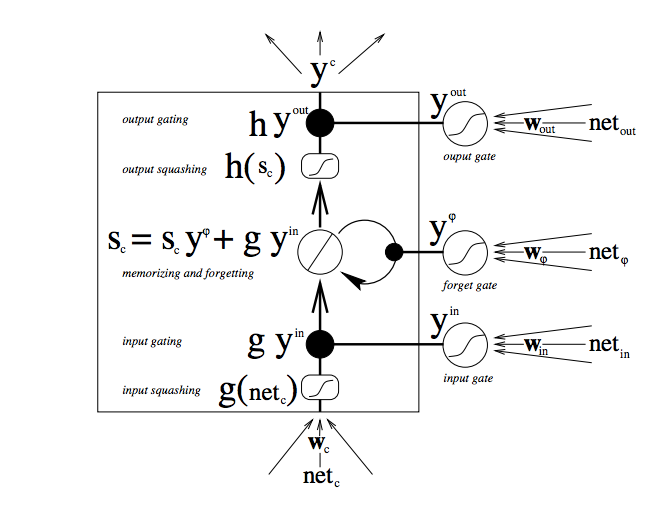


图38 DL4J中block的处理

首先时cell的输入g()和输入门通过乘法得到cell的输入。其次通过cell前一个状态和激活门的值进行乘法得到,接着更新cell的状态，之后根据输出门和cell的状态得到块的输出。

下面使用Christopher Olah的《Understanding LSTM Networks》中对该行为的解释，更加简单明了，由于3个门和一个输入的值一样，下面是Christopher Olah的理解。

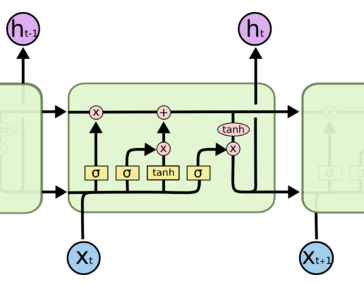


图39 Christopher Olah关于block的理解

从左至右依次是遗忘门、输入门、输入、输出门。

对于遗忘门：

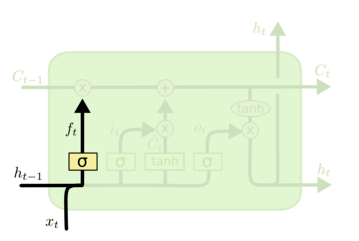


图40 Christopher Olah关于block的理解

对于输入：

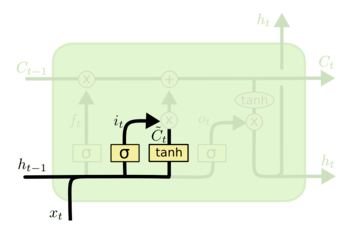


图41 Christopher Olah关于block的理解

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：

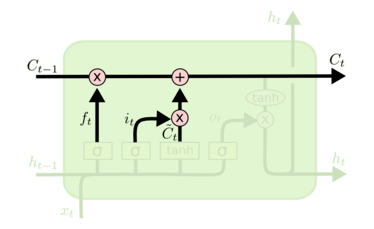


图42 Christopher Olah关于block的理解

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，如下示：

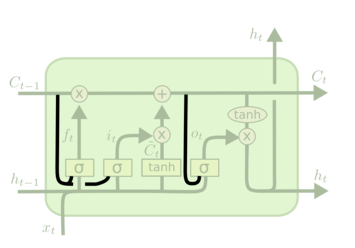


图43 Christopher Olah关于block的理解

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）：

在此我们以及介绍了LSTM的基本组成单元，除了这个单元以外，其和普通的循环神经网络没有什么区别，如下示：

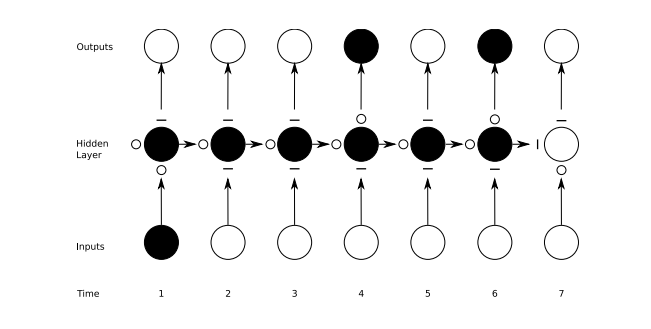


图44 LSTM网络（DL4J）

门的运作方式，其中横线代表关闭的门，而空心的小圆圈则代表打开的门。在隐藏层下方水平一行的横线和圆圈就是遗忘门。前馈网络只能将一个输入映射至一个输出，而循环网络则可以像上图那样将一个输入映射至多个输出（从一张图像到标题中的许多词），也可以进行多对多（翻译）或多对一（语音分类）的映射。相比前馈神经网络，我们还需要设置序列的长度，由于上面我们只介绍了简单的LSTM和一个更新的LSTM，但如今有许多变体，因此我们需要设置Cell,初始化类型等。

**3.5 小结**

本章主要介绍了处理回归问题的几种方法以及一种优化算法，在理论上说明解决该问题的方法的多样性。并在后面回归系统的建立提供理论基础，同时让我们了解到机器学习这个大家族的几种流行的方法，从最早的支持向量回归，到BP神经网络，再到循环神经网络。尝试了很多方法，但其中原理和优化算法等有很多很多中，如果在这段时间里每一个都去研究也是不可能的，因此我选择了SVR、BP神经网络、以及LSTM。首先SVR在回归处理上有一定的很长的历史和令人满足的准确率，而BP神经网络，由于神经网络的兴起和BP神经网络的相对简单，因此也选择使用了，而LSTM在序列处理上有很多的有点，因此也选择了。而人工网络在处理上，存在很多方法，自己可能也没有理解透彻，只是简单的运用了一下。

**第4章 基于多源数据的高速公路流量预测系统的设计**

我国高速公路的不断发展，行车在高速公路上变得很常见，然而给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、对高速公路损坏部分的修复。虽然看起来没有什么问题，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，即根据历史的交通流量和周边环境的数据对未来交通流量进行预测。

**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，然后可以先在验证集上跑一遍进行验证，再可以进行预测。

**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。

3、预测模块，即将数据导入训练好的模型中，根据不同模型得到的值，进行一种组合。

**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，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，处理训练就是调参，因此训练过程事关重要。接着就是测试，最后就是预测，测试是判别我们训练的样本是否符合我们的要求，我们使用数据集中没有参数训练的样本进行测试，最后才根据我们当前的数据进行预测。

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。

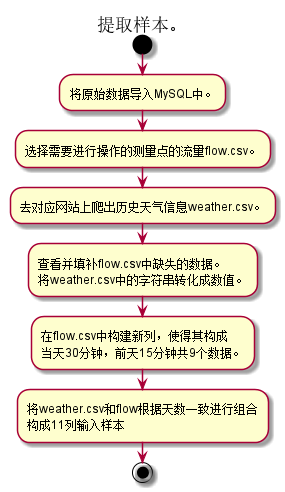


图45 提取数据样本的流程图

对机器学习算法的训练过程，将提取的数据进行分类，分成训练集和测试集，然后在多个机器学习的算法中进行训练，最好组合几种算法进行回归：

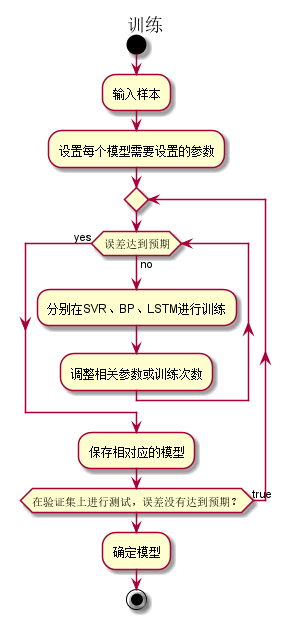


图46 对数据进行训练的流程图

对训练好的模型进行预测的流程过程：

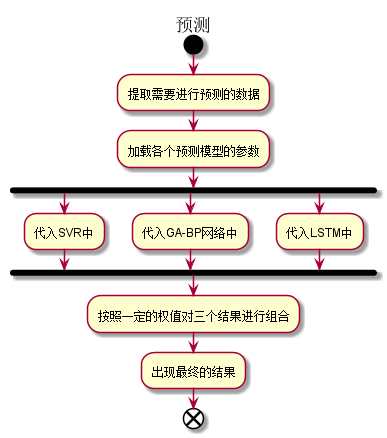


图47 使用模型进行预测的流程图

**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个列的逗号分隔符文件（.csv），我们使用Python中的文件处理open对我们的数据进行操作，读到每一行的数据时，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列作为特征(X)，为了对模型进行更加好的选择，我们需要定义训练集和测试集。训练集是用来训练我们的预测模型的，而测试集是用来保证我们训练的模型具有一定的泛化能力，避免其过拟合。

传统划分数据集的方法：留出法、交叉验证法、自助法。一般都是将训练集和测试集进行七三分，70%的数据进行训练，30%的数据进行测试。留出法：直接将数据集的70%拿出来作为训练集，30%的数据进行测试集。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。

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，我们使用零-均值规范化，将数据处理成符合标准正太分布。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。

因此最终我们的数据都是符合标准正太分布，并且打乱了样本顺序的数据集，并且实现了训练集和测试集分开的操作。

**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SVR、GA-BP（用GA优化BP神经网络的初始的权值）、长短时间记忆模型LSTM，即一种循环神经网络。

对于支持向量回归而言，我们设定的参数有C，，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，遗传的代数，变异的概率，交叉的概率。以及BP神经网络的层数和每层的神经元个数，权值更新的大小，学习率。对于LSTM，我们需要确定序列的大小。

**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。

解释方差(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：

解释方差的最大值是1.0，表达式的值越接近1.0代表预测的结果越好，反之代表结果越差。

平均绝对误差（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：

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。

平均平方误差（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：

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：

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。那么该方法的计算过程如下示：

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，我们需要一个选择训练数据集的样本的按钮，点击按钮，我们选择我们需要的数据，训练完之后，我们需要知道其误差是多少并且保存其模型参数，之后我们需要对该模型进行测试，我们需要了解一下真实值和预测值之间的差距有多少，即我们模型的效果如何。最好就是我们的预测功能，点击该按钮我们就能得到我们需要的值。

**4.5 小结**

本章主要对基于多源数据的高速公路流量预测系统做了详细的设计，首先对系统的需求进行了分析，并提出了系统的总体设计。然后对系统功能进行了分析讨论。对系统进行了详细设计。由于自己是在尝试中进行开发得，可以说成式增量式的吧，对于这一块，也只是做完之后写的，然后自己在实际过程中，不知道设计和实现的具体分界点，因此所谓的设计只是最终成功的一种对应结构，从一开始的对交通流量数据的处理、天气数据的爬取。到后来模型的选择，学习算法的选择、优化算法的选择、参数的选择等等。其中也有很多弯路。但自己始终坚持自己的毕业设计的目标---高速公路短时流量预测，在软件工程里面应该算的上是基线。因此我的觉得，在做不是那么大项目的时候，各种设计啥的我们可以简单化，主要抓住我们最初的基线进行开发就可以了。

**第5章 基于多源数据的高速公路流量预测系统的实现**

**5.1 系统开发环境**

* 操作系统：Ubuntu16.04
* 集成开发环境：Pycharm、TensorFlow1.0、sklearn。
* 编程语言：Python3.6

**5.2 系统实现**

**5.2.1 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，根据其官网提供的文档显示，其数据读取一般有三种方式。分别是供给数据（Feeding）、从文件中读取数据、预加载数据；另外就是传统的数据读写。

由于预加载数据，是通过在图中定义变量和常量来保存数据，一般只适合少数数据，因此就没有使用该方法，而在此系统中，我使用框架提供的从文件中读取再给图供给。同时也尝试直接读取文件，读取每一行，将其存取在numpy的数据中。之后提供给图中的向量进行训练。对于两种不同的方式，其第二阶段是一致的，即都是给向量进行“赋值”操作。因此两者的不同点就在于读取文件的方式。

**5.2.2 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），然后就是拉格朗日乘子，对偶问题进行解决。而在此系统中，我除了使用TensorFlow框架，而调用的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。再就是将数据喂给这个SVR模型。而此时我使用的系统默认的参数，C=1.0，松弛因子为0.5。

对于BP神经网络。首先设置网络的结构。主要是隐藏层的结构，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，对于根据算法，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。

对于循环网络中的长短时记忆（LSTM）。我选择了学习率为0.0006，序列长度20，隐藏black10个，使用的cell是基本的tf.nn.rnn\_cell.BasicLSTMCell()，初始化其状态是cell.zero\_state()等。

**5.2.3 模型验证功能实现**

对于验证，即测试，我们把测试集在之前训练好的模型跑，看每个模型的泛化能力，在这里我们使用的是第四章介绍的模型评估方式，对于每一种方法我们都需要进行计算，然而在sklearn库中，我们有评估模型的方法。在sklearn.metrics下的explained\_variance\_score、mean\_absolute\_error、mean\_squared\_error、r2\_score。对于每一个方法，其参数形式为model(y, )就可以了，在命令行中，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y和测试集代入sklearn.metrics下的评估模型中。

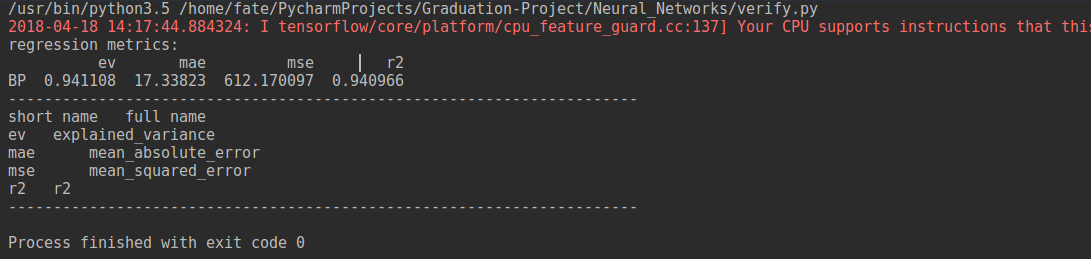


图48 评估模型

**5.2.4 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。

**5.2.5 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。

**5.3 系统展示**

**5.3.1 系统界面展示**



图49 界面设计图

**5.3.2 选择文件**

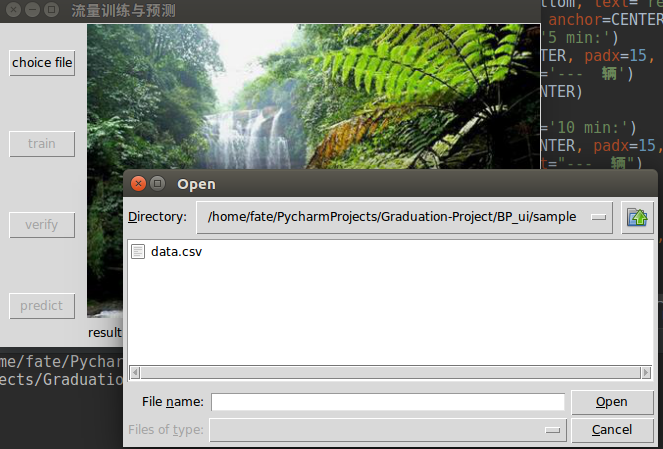


图50 选择文件的界面

**5.3.3 模型训练**

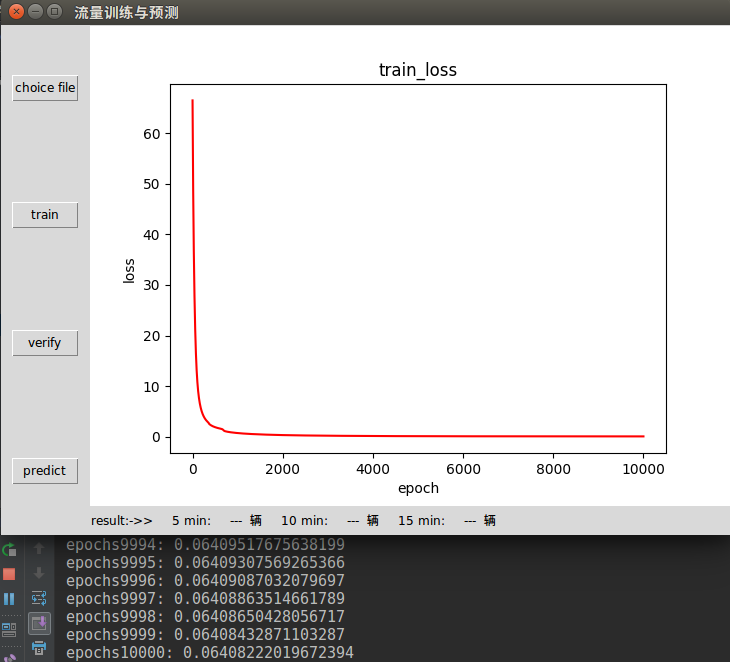


图51 训练之后的界面

**5.3.4 模型验证**

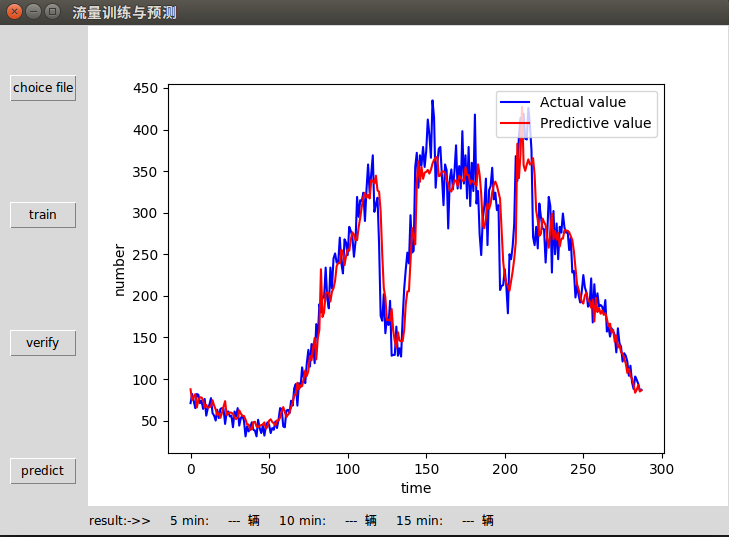


图52 模型验证之后的界面

**5.3.5 进行预测**



图52 预测之后的界面

**5.4 小结**

本章主要介绍了基于多源数据的高速公路流量预测的相似流域查找系统的实现过程，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。而对于支持向量回归、长短时间记忆等的研究更加浅薄。同时也有很多自己能力处理不了的不足。

**第6章 总结与展望**

毕业论文从选题到开始实施，到最后的完结，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、无力的浪费，同时堵车造成的环境污染问题也不容忽视。虽然目前的电动汽车的发展，但其还是没有普及化。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**

在本次毕业设计过程中，我学习了很多之前没有接触过的技术，并在实践过程中逐渐掌握了这些技术，为我以后的学习打下了坚实的基础。

我首先要感谢我的论文指导老师——谢在鹏老师悉心的指导。在每周的毕设讨论会议上谢老师都会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。在论文写作期间，谢老师都会对我进行督促。谢老师对待工作严谨的态度、渊博的学识、耐心的教诲让我受益匪浅。谨此表示衷心的感谢！

然后还要感谢实验室里面的师哥，他们在技术上给予我的指导和论文写作上的给我的建议让我的毕业设计进展事半功倍。

最后，请再次让我感谢所有在毕业设计中给予我帮助的人，使得我的毕业设计能顺利如期完成。

参考文献