

文本复制检测报告单(全文标明引文)

№:ADBD2018R_2017031309522720180308141956835336383360

检测时间:2018-03-08 14:19:56

检测文献: 60580031136815381_城市交通路口短时流量预测2

作者: 城市交通路口短时流量预测

检测范围: 中国学术期刊网络出版总库

中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库

中国重要会议论文全文数据库

中国重要报纸全文数据库

中国专利全文数据库

互联网资源(包含贴吧等论坛资源)

英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)

港澳台学术文献库

优先出版文献库

互联网文档资源

图书资源

CNKI大成编客-原创作品库

个人比对库

时间范围: 1900-01-01至2018-03-08

检测结果

总文字复制比: **8.2%**

跨语言检测结果: **0%**

去除引用文献复制比: **8.2%**

去除本人已发表文献复制比: **8.2%**

单篇最大文字复制比: **1.8%** (城市交通路口短时流量预测)

重复字数: [951]

总段落数: [2]

总字数: [11556]

疑似段落数: [2]

单篇最大重复字数: [210]

前部重合字数: [0]

疑似段落最大重合字数: [553]

后部重合字数: [951]

疑似段落最小重合字数: [398]



指标: ☐ 疑似剽窃观点 ☒ 疑似剽窃文字表述 ☐ 疑似自我剽窃 ☐ 一稿多投 ☐ 疑似整体剽窃 ☐ 过度引用 ☐ 重复发表

表格: 0 脚注与尾注: 0

4.5% (398) 60580031136815381_城市交通路口短时流量预测2_第1部分 (总8872字)

20.6% (553) 60580031136815381_城市交通路口短时流量预测2_第2部分 (总2684字)



(注释: 无问题部分 文字复制比部分 引用部分)

1. 60580031136815381_城市交通路口短时流量预测2_第1部分

总字数: 8872

相似文献列表 文字复制比: 4.5%(398) 疑似剽窃观点: (0)

| | | |
|---|--|----------------------|
| 1 | 基于支持向量机的变形预测方法研究 程功(导师: 罗志清) - 《昆明理工大学硕士论文》 - 2010-05-01 | 1.1% (95) 是否引证: 否 |
| 2 | 新型干法水泥生产窑尾分解率软测量研究 刘慧(导师: 袁铸钢) - 《济南大学硕士论文》 - 2010-05-20 | 0.8% (71) 是否引证: 否 |
| 3 | 脑活动状态EEG信号解码方法及其应用 李俊华(导师: 张丽清) - 《上海交通大学博士论文》 - 2012-09-01 | 0.8% (68) 是否引证: 否 |
| 4 | 支持向量机在水力发电机组轴心轨迹自动识别中的应用研究 彭兵;程建;曹长冲;尹永珍;司汉松; - 《水电与新能源》 - 2016-02-28 | 0.6% (49) 是否引证: 否 |
| 5 | 柔性薄板装配不等间距定形焊点布置优化 | 0.5% (45) |

| | | |
|---|---|-----------------------|
| 曲兴田;马强;李笑男;王昕;郑海生;吴晟宇; - 《计算机集成制造系统》 - 2015-12-14 0 | | 是否引证：否 |
| 6 | 基于SVM的公安情报自动分类系统的设计与实现 牛琳(导师：司志刚) - 《解放军信息工程大学硕士论文》 - 2007-04-20 | 0.4% (36) 是否引证：否 |
| 7 | 基于测距信息的室内机器人定位技术研究 郑向梅(导师：杨庆芬) - 《石家庄铁道大学硕士论文》 - 2017-06-01 | 0.4% (32) 是否引证：否 |
| 8 | 手势建模算法研究及其应用 魏朝龙(导师：滕少华) - 《广东工业大学硕士论文》 - 2011-05-01 | 0.3% (30) 是否引证：否 |
| 9 | 城市交通路口短时流量预测 张金飞;黎英;高伟;黄名钊; - 《软件》 - 2018-01-15 | 0.3% (29) 是否引证：否 |

原文内容 红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

交通流数据分析与预处理

1交通流基本理论

道路交通主要是由道路上形形色色连续流动的人流和车流所构成的，行人以及车辆在路面上不断行驶就会呈现一定的特征出来，交通流就是由车流和人流所共同形成的。然而在本文中研究的“主人公”仅仅是车辆形成的车流，因此在文章的后续研究中都是将车流作为交通流量，交通流状态就是指连续不断的交通流运行所表现出来的特性。

不同类型的车辆、不同的旅游目的地以及不同的出行目的都会促使交通流发生不同程度的改变；而且天气条件、道路状况、交通情况都会对交通流产生很大改变，所以根本不好用很清晰恰当的物理量来表达这都是因为交通流的混乱特性造成的。

虽然是有这些情况存在，经过不断地研究分析通过大量的实验观察，还是会发现在一条件下交通流的状态变化也并不是无迹可寻还是有一定的规律存在。实际情况中我们是可以用我们所熟知的交通流状态的特征参数去表示存在的这些规律和交通特征。对于交通流参数我们是可以根据研究和观察分析的不同对象分别从宏观和微观这两个方面来对其进行表示。在研究对象仅仅就是单一的车辆主体的时候，我们使用微观的参数进行表示，微观参数包括像车头间距以及车头时距等。而当我们把交通流这样个整体作为我们的研究分析对象的时候，肯定是从宏观的角度用宏观相关参数进行表示，这些参数是我们比较熟悉的例如：速度、密度、交通流量、占有率及排队长度等，这当中都是将密度、速度、交通流量作为交通流三要素。

在本文后面的实验分析当中都是从宏观的层面来进行研究分析的，在这章下面主要介绍流量、速度、占有率这三个基本参数。

(1) 流量---也叫交通量。它指的是在一个时间单位内通过某个路面断面的车辆数量，而且会由于时间以及地点的不一样变化也会非常大。短时交通流通常指的是在15min内道路断面上通行的车辆总数。在本文的研究中选取的对象即为流量，选取的时间间隔设定在10min,也就是每隔10min道路上某个断面路口总共通过的车辆数目。

(2) 速度。它指的是研究对象车辆在一个时间单位内所行驶过的距离。事实上的交通流系统结构纷繁复杂，这其中不但包含许许多多的车辆，而且不同的车辆之间的特征差异也是很大的，故而我们所说的交通流的速度其实是指所有车辆的一个平均速度。时间平均速度（Time Mean Speed，TMS）以及区间平均速度（Space Mean Speed，SMS）都是包括在交通流的平均速度之中。

在实验的观测时间段内，路过所在实验路段断面内的所有车辆的瞬时速度的平均值即为时间平均速度，详细计算公式如（3-1）所示：

式（3-1）中， v_t 即为时间平均速度， n 代表实验时间道路断面内观测到的所有车辆数目。

在实验进行的观察时间区间里面，所观察道路路段的长度与行驶过此路段的所有车辆平均行驶时间的一个比值即为区间平均速度，详细计算公式如（3-2）所示：

式（3-2）中， v_s 即为区间平均速度， L 代表观察路段的长度， n 观察路段断面上行驶的车辆数目， t_i 表示的是第 i 辆车辆行驶通过观察路段所花费的时间。

(3) 密度。从理论上讲，密度指的是在单位长度上的道路路面中瞬间行驶过的机动车的数量。不过事实上，在实际情形下很难再瞬间去测量通过的车辆数目，所以退而求其次，就用车辆占有率来取代密度表示，占有率包括时间占有率和空间占有率。

空间占有率表示的是在某一时间段内，直白的说就是观察的道路断面上行驶过的所有机动车车辆的长度总和除上这段断面的长度得到的值。空间占有率可以更清楚的显示道路断面上的占用状况，事实上要想求得所有机动车的长度总和还是个比较困难的事情。空间占有率的详细计算公式如（3-3）所示：

式（3-3）中， k_s 即为空间占有率， L 表示所观察的道路断面距离长度， n 表示观察路段上行驶过的机动车车辆总数， l_i 代表的是第 i 辆车的长度。

时间占有率通常指的是机动车辆在某一段时间中所占有车辆检测器的时间总和与用于检测的总时间的一个比值。如果检测断面上的机动车辆行驶的比较快则占用的检测器时间必然会很少，那么时间占有率就会相对小，这样就代表此路段不会拥堵；如果断面上的车流量增加，很显然检测器就会被机动车辆占有很长的时间，则相应的时间占有率定会变大，此时代表着此路段比较拥堵。时间占有率的详细计算公式如（2-4）所示：

式（3-4）中， k_t 即为时间占有率， T 表示检测器的检测总时间， n 代表检测时段内行驶过的机动车辆总数目， t_i 代表着第 i 辆车行驶过检测器的时间大小。

3.2实验数据来源

论文的仿真数据来源于交通运输部出行云中的合肥市示范区黄科路口相关数据。本文选取的数据为其中的微波检测数据。微波检测数据表名为DT_LANE_REPORT_H，包括编号、设备类型、设备编号、采集时间、时间占有率等多个字段信息，选取的时间为2016年6月30日至2016年7月1日共两天的历史数据，数据详情请见下表。

3.3交通流数据预处理

在本文实验中所采用的交通数据源最初是通过微波检测器收集而来的，而事实上在真正的工程应用里面，会因为例如交通数据采集设备的一些固有缺陷或者是在数据传输过程中传输设备发生的故障以及其他外界因素的干扰等等不同的原因，往往就导致了这些检测器收集到的原始交通数据不是有很高的精度问题。但是在这些检测器上收集到的原始的交通原数据是进行短时交通预测模型里面的一个数据基础，因此它对这个模型的可信度和有效度是直接产生了不可低估的影响作用。这样的话对于从检测器收集到的第一手交通数据存在着质量问题是不能直接放到模型中使用的。那么既然如此，理所当然的必须对从检测器收集到的交通数据进行必要手段的预处理工作，将预处理之后的交通数据才能放到最后的模型进行试验。所以，对采集到的交通数据进行预处理工作是进行建模预测的第一步，这一步必不可少，是整个过程中很重要的一环。

3.3.1异常数据修正

在获取到的交通数据当中必不可免的会有一些异常数据，这些异常数据的来源可能是由于检测器在某个阶段产生了故障或者失灵也可能是在数据传输的过程当中碰巧遇到设备的故障等等原因。异常数据可以将其简单分为数据缺失和数据错误两类。

(1) 对异常数据中的数据缺失的处理方法。

本文使用的数据来源是通过微波检测器采集得到的，它的采集时间间隔也是固定的，所以本文的数据采集间隔为1min，这样从理论上讲，一天24小时会采集到1440条交通流数据，然而实际得到的交通流数据因为数据的赘余或者缺失等原因，一个检测器在一天收集到的真正的数据量是围绕在1440这个数字上下波动的。对解决数据缺失的常用到的方法像有加权平均、历史平均法等等。在本文使用到的处理方法是加权平均，具体实施步骤如下：

第一步：拿到缺失数据前一天而且是同一时间点的历史数据 q_{k-1t} ；

第二步：拿到缺失数据在当前时间点的前一时刻的实测值 q_{kt-1} ，然后将 q_{k-1t} 、 q_{kt-1} 使用加权平均法，即可恢复得到缺失数据，详细公式如(3-5)：

式(3-5)中， q_{kt} 代表第 k 天 t 时间点的修复得到的修正值， ω 代表加权因子，表示

$t-1$ 时间点的实际数据及历史数据在丢失时间点数据进行修正的时候所拥有的影响的比重大小，第 k 天 $t-1$ 时间点的实际数值用 q_{kt-1} 表示，第 $k-1$ 天 t 时间点的历史数据由 q_{k-1t} 表示。

(2) 对交通数据中产生错误数据的处理方法。

理论上来说，我们是可以进行定义一个合适的阈值，通过这个阈值来筛选交通流数据中的错误数据，因为在一定的时间区间内的各个交通流参数的取值会分布在一个合理的范围中，那么就可以把落在阈值范围外的数据定性为错误数据。而且我们在已有的交通流理论和大量的实践基础上可以给这些交通流参数的大概取值范围给确定下来，流量、速度、占有率这些交通流参数的合适阈值范围设定如下：

a.流量

通过交通路口上的微波检测器收集到的交通数据中交通流量 q 折算后的合理范围如式(3-6)：

C —代表道路路段上每小时所能通行的机动车数即路段的通行能力(veh/h)；

T —代表微波检测器收集交通数据的时间间隔，本文中为1min；

B —代表对交通流量的一个修正权重，范围设定在1.3~1.5。

b.平均速度

平均速度 v 也是通过路口微波检测器采集所得，范围设定如式

dv —代表设计速度，本文中取80km/h；

vf —代表修正的权重值，范围设定在1.3~1.5。

c.占有率

在路口中微波检测器收集到的占有率数据为时间占有率 Occ ，其范围设定如式

3.4交通数据归一化

在本文中，所做的归一化其实就是将通过微波检测器采集到的交通数据按照一定的计算标准，使用某种具体的方法将其规范化到一个合适的范围中。进行归一化的目的在于得到的交通数据参数不唯一，数据量大，数值差异也比较大，将其统一到一个合理的范围中去后使得后续的运算更加方便和规整。

为了能很好地去除由于数据的太大差异性给实验带来预测结果的干扰，而且本身交通数据的变化幅度就比较大，因此在实验之前必须对得到的交通数据进行一个归一化的处理操作，且归一化后的数据范围映射在[0,1]之间。

本文的主要实验操作都是在MATLAB2015a基础上进行的，此环境本身也自带了多种进行归一化的函数方法，本文在实验中利用了常用的mapminmax函数方法对数据进行归一化的操作。Mapminmax方法函数的具体使用格式如式

式(3-9)中， Y 表示的即为归一化操作后得到的数据矩阵， Z 代表的则是记录了相关信息的一个结构体， X 表示的即为需要归一化的初始数据矩阵。

对某一个交通数据 $Data_i$ 的具体计算原理如式

3.5本章小结

本章首先阐述了关于交通流的基础知识，并且表明本文实验的短时交通流量预测的研究对象为交通流，同时重点介绍了三个基本参数：流量、速度、占有率；其次交代了本文的实验数据来源为交通运输部出行云中的合肥市示范区黄科路口相关数据；最后主要交代了对获取的交通数据的分析和预处理，分别有异常数据的处理方法、错误数据的处理方法和数据最终的归一化。这章的知识主要为本文后面的算法模型搭建提供一个数据基础。

第4章基于支持向量回归机的短时交通流预测

学习和模仿能力在我们人类的整个繁衍、进化和发展的历程当中毫无疑问是我们可以生活在今天这样一个美好时代的很重要的原因。我们通过不断地对信息的收集、研究、分析和归纳来找到包含在数据背后的规律，有了规律就可以顺藤摸瓜通过这些规律来进行预测，从而发现我们原本不能直接观测得到的信息。渐渐地，我们希望计算机也可以像人的思维一样来模仿我们的这种学习思维方式，这就产生了机器如何学习的问题即机器学习的由来。这几年机器学习的一个研究大热点就是AI即我们所说的人工智能，它的一个重要的意义在于希望通过计算机来发现海量数据中存在的相关性，找出可用的有价值的信息，基于此我们进一步的来进行判断、分析和研究去预测未来的数据信息也就是“未卜先知”的能力。

这些年机器学习的快速发展得益于统计学为其提供了丰厚的理论支持和基础。但是传统的统计学和本文所要说到统计学习理论 (Statistical Learning Theory, STL) 也是不尽相同，在本文中其实主要是针对在没有丰富的样本数据的情况之下来对机器学习开展的研究。支持向量机 (Support Vector Machine, SVM) 的理论思想是通过某种方法将不属于同一类别的样本进行正确的归类，SVM就属于是本文所用的统计学理论下的一个完美产物。

支持向量回归机 (Support Vector Regression, SVR) 其实是借用了支持向量机的思想，将分类的思想上升到回归的问题，而且也在很多工程实践中得到了很好地应用。尤其在非线性的情况，数据具有多维特征，SVR算法将数据映射到高维空间，来解决在原始空间中线性不可分的问题。这是由于SVR中核函数的优良品质，这样的非线性扩展在计算量上并没有比原来复杂多少，这一点是非常难得的。在短时交通中，交通流具有规律性弱、不确定性强等非线性的问题，而SVR本身在解决小样本、非线性等问题有着很好的优势，且SVR结构不是很复杂，易实现，不会陷入局部解，所以比较契合在短时交通上的工程应用。

4.1统计学习理论

4.1.1学习问题的描述

统计学习理论是建立在一整套丰富的机器学习理论之上的，具有很扎实的理论。相比较于传统的统计学，STL可以在没有很充足的实验数据的情形下去分析机器学习的规律，这就提供了一个可以在原本没有丰富样本的情况下进行机器学习的统一参考的框架。

对机器学习问题其实就是在式 (4-1) 的基础上进行如下研究：

假设有 n 个独立同分布的数据样本：

通过对数据样本的研究分析，期望能够在 $\{f(x, w)\}$ 函数里面，求解得出一个最优的 $f(x, w_0)$ 可以使期望风险 $R(w)$ 即式 (4-2) 最小。

式 (4-2) 中， $f(x, w)$ 代表的是预测函数集， w 表示广义上的参数，所有的函数集合都可以用 $\{f(x, w)\}$ 表示；通过 $f(x, w)$ 对 y 预测所产生的损失为 $L(y, f(x, w))$ ，损失函数会由于不同类型的学习问题有多种形式，像模式识别类型的问题损失函数的定义如式 (4-3)：

在对连续的回归估计问题上，损失函数如式 (4-4)：

对于概率密度的估计问题，密度函数用 $p(x, w)$ 表示，对损失函数的定义如式 (4-5)：

3.1.2经验风险最小化

使得期望风险最小才是我们进行机器学习的目的，然而我们手中能够利用的数据信息只有如样本数据 (4-1) 这般，这样我们希望式 (4-2) 的期望风险就没有办法进行计算，因此在传统的研究方法中就另辟蹊径，将样本来定义出一个经验风险，采用的是经验风险最小 (Experience Risk Minimization, ERM) 准则。

构造出来学习算法使得对式 (4-2) 的估计即期望风险的值能最小，式 (4-3) 的损失函数即经验风险对于模式识别类型的问题来说就是得出训练数据样本的错误率；式 (4-4) 的损失函数即经验风险对于函数回归估计问题来说就等同于训练误差；式 (4-5) 的损失函数即经验风险对于概率密度估计的问题可以视为最大似然函数方法。理论上来说，把期望风险最小化用ERM准则来进行替代也没有很充足的理论支持，仅仅是从直观上理解我们认为这是合理的。例如，若果当 n 的取值倾向于一个无穷大的时候，此时式 (4-6) 的结果就会和式 (4-2) 相差无几了，因此，对于许多的问题，他们的数据样本其实是有限的，那么在这样有限的数据样本下通过ERM准则得出的最小经验风险就未必是真的最小。

4.1.3结构风险最小化

基于VC维的理论基础，统计学习理论中推广性界分析探究了不同类型的方法函数集以及真实风险和经验风险的关系。实际的风险 $R(w)$ 与经验风险 $R_{emp}(w)$ 两者存在至少 $1-\eta$ 的数学概率并满足式 (4-7)、(4-8) 两个关系式，这是对二分类问题总结得出的结论。

式 (4-7)、(4-8) 中， n 代表的是数据样本的数目， h 表示方法函数集合的VC维。

可以得出若 h/l 比较大的话，会使得学习机置信区间变大，那么就会使得真实风险增加进而会导致较大的误差产生。相反，若数据样本量很大， h/l 会变得比较小，学习机的置信区间会缩小，此时ERM的值与实际值就会很接近。

若我们的数据样本是一个定值时，学习机的复杂度如果越小即VC维越小，则其置信范围必然会减小，这样检验风险与实际值的差异会缩小。那么当需要建立分类器就需要考虑不仅要尽可能小的经验风险还得把控好VC维，就可以减小置信范围。

最小化结构风险的思想：若想要达到一个最小化的结构经验风险，方法就是让式（4-7）中的 $R_{emp}(w)$ 和 $\gamma(h/n)$ 进行制约，一起趋向最小，而且得使置信区间达到最小也即是让 h 的值要尽可能的小。

对于公式（4-7），当所需的训练样本数据量 n 有限的情况下，则可以通过 $R_{emp}(w)$ 和 h 两个参数来调整结构风险 $R(w)$ ：

（1）机器学习的ERM由函数集合 $f(w, x)$ 决定的，利用 w 可以对经验风险进行掌控。

（2）学习机的函数方法集合会对VC维的维度产生影响，首先对函数集合进行结构从而搭建出函数子结构与维度 h 关联，把VC维与函数集合联系在一起。

对方法函数集 $\{f(w, \alpha), w \in \Gamma\}$ （ Γ 代表的是抽象参数集合）进行结构化采取的方法如下：

对函数集合 $S_1, S_2, S_3, S_4, S_5, \dots, S_n, \dots, S$ ，当中 $S_k = \{f(w, \alpha), \alpha \in \Gamma_k\}$ ，对任一函数集 S 里面的元素 S_k 都会存在一个VC维 h_k 且是有限的，并有 $h_1 \leq h_2 \leq h_3 \leq \dots \leq h_k \leq \dots \leq h$ 。对一组数据样本 $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_l, y_l)$ ，在函数集 S_k 里面利用结构风险最小化的思想选取出一个方法函数 $f(w, \alpha_k)$ 以期得到ERM，并保证 S_k 有最小的置信范围。

4.2 支持向量机理论

4.2.1 支持向量机原理

SVM算法最初是用在解决线性可分的问题上并基于此慢慢发展而来，其理论思想通过图4-2展示出来的二维平面进行理解。按图中所示有两种类型的数据样本存在，中间的一条为分类线 H ， H_1 、 H_2 是和 H 这条分类线平行的，而且这两条是与 H 距离最近的平行线， H_1 与 H_2 之间的距离称为分类间隔。当这个分类线 H 可以将两类样本数据正确分离并且分类间隔最大的时候此时的分类线为最优。

令分类线为 $(w \cdot x) + b = 0$ 并对其做标准化，对能够进行线性可分的数据样本集合 $S = \{(x_i, y_i)\}$ 使其满足式（4-9）：

能够使分类间隔 $\|w\| / 2$ 的值最小或者是 $2 / \|w\|$ 的值最大同时也可以将训练数据样本 (x_i, y_i) 进行正确的分到所属的类别的这种分类面被称之为最优分类面。SVM中的支持向量就是位于 H_1 、 H_2 线上的数据样本点。支持向量机的核心思想在于要实现主要的泛化能力，这是要使得分类间隔最大化去实现的。

所以，通过线性可分条件来建立最优超平面将求最大化分类间隔问题转为二次规划为题：

通过对式（4-10）采用Lagrange乘子法进行求解，可以将二次规划问题再次转为解Lagrange函数鞍点的问题：

式（4-11）中， α_i 表示拉格朗日乘数且 $\alpha_i \geq 0$ 。

据KKT条件若要取得最优解必须满足式（4-12）：

对式（4-11）进行求解，有唯一的解。并将 α_i^* 设定为得出的最优解，那么可以得出：

在公式（4-13）中， α_i^* 即代表所需要的支持向量。 b^* 表示分类间隔阈值，它可以利用上式（4-12）得出。最终可以得出需要的最优分类面：

以上对最优分类面的求解是对线性可分的情况下进行的，不能用于非线性的问题，若需要解决非线性的问题，需要引入核方法，可以将非线性下的问题由低维空间反映到高维的特征空间，然后同样的即可解出最优分类面。基于泛函数的理论，若核函数 $K(x_i, x_j)$ 能够达到Mercer定理的要求，那么就有某一变换空间的点与其点积对应。因此，想在引入核函数 $K(x_i, x_j)$ 后不增加计算复杂度成功的解决非线性转变成线性后的分类，需要着重考虑选取理想的核函数，此时最优的分类函数方法为：

4.2.2 支持向量回归机原理

SVR算法主要可以用来解决线性和非线性方面的回归问题，眼下SVR存在两种分别是 ϵ -SVR和 ν -SVR。文章使用的是 ϵ -SVR进行短时交通的预测。 ϵ -SVR的理论思想为：在拥有 m 个样本数据 $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ ， x_i 代表输入样本， y_i 代表输出样本。首先，对原始的输入样本数据利用非线性变换将其投影到高维的空间中，将非线性回归变成线性回归。则在投影到的高维特征空间中，我们可以建立一个最优超平面完成分类。

在公式（4-16）， b 代表了偏移量。输入数据样本经过一个非线性变换的映射变换到高维的特征空间中去之后，达到了将非线性回归问题转为线性回归问题的求解目的，将不敏感损失函数定义为式（4-17）：

SVR问题目的就是企图找到一个合理的 $f(\cdot)$ 能够使得 $E(w)$ 取得一个最小值：

在式（4-18），线性权重用 $w = (w_1, w_2, \dots, w_m)^T$ 表示； C 代表惩罚参数，值过大会造成过拟合，值太小分类性能就会很差。

在对问题的求解的时候我们引入了两个非负的松弛变量 ξ_i 和 ζ_i ，于是得到（4-18）等价的对偶问题：

上式中， β_i 和 $\bar{\beta}_i$ ，分别代表的是 ξ_i 和 ζ_i ，这两个引入的松弛变量所对应的拉格朗日乘子， $K(x_i, x_j)$ 即为合适的达到Mercer定理要求的核函数。

因此对于测试样本数据 x 所求的输出可按是（4-20）进行预测：

从上式可以得出，支持向量回归机与神经网络在形式上比较相像，任一中间节点都有一个支持向量与之对应，输出的都是对一个中间节点的线性组合，如下图所示：

在SVR的推导计算后，最后SVR的分类函数是通过将输入的未知的数据向量与每个支持向量的内积，因此对SVR回归计算的复杂度仅仅是由支持向量的个数来决定的。

4.2.3 松弛变量

在解决实际的样本数据中，会有各种原因产生使得很多时候不是所有的问题都可以像线性可分那样简单，在一些情况下会

有不能进行线性分离的点出现，这些数据点可以称之为“离群点”。

对数据样本中，我们将间隔最小的点设定其间隔为1，同时加入新的变量为松弛变量 ξ ，则式（4-9）可改写为如下：

公式（4-21）中，引入的变量 $\xi \geq 0$ ，通过约束条件进行计算是允许其小于1的，然而若有一数据点间隔小于1的话，实际上是很难被准确分类，则目标函数可写成式（4-22）：

对（4-22）式中，C即为惩罚参数，其值的变化可以用于减轻错误分类样本的惩罚； ξ_i 表示分类超平面训练数据的偏差值，若数据样本点即为离群点的时候，有 $\xi > 0$ ，相对应的非离群点有 $\xi = 0$ 。

| 指 标 | | |
|--|---|------------------------|
| 疑似剽窃文字表述 | | |
| 1. 数据映射到高维空间，来解决在原始空间中线性不可分的问题。这是由于SVR中核函数的优良品质，这样的非线性扩展在计算量上并没有比原来复杂多少， | | |
| 2. 60580031136815381_城市交通路口短时流量预测2_第2部分 | | 总字数：2684 |
| 相似文献列表 文字复制比：20.6%(553) 疑似剽窃观点：(0) | | |
| 1 | 城市交通路口短时流量预测 张金飞;黎英;高伟;黄名钊; - 《软件》 - 2018-01-15 | 6.7% (181) 是否引证：否 |
| 2 | 基于支持向量回归的短时交通流预测方法研究与应用 武琼(导师：王夏黎) - 《长安大学硕士论文》 - 2016-04-28 | 6.1% (164) 是否引证：否 |
| 3 | Hadoop环境下基于SVR的短时交通流预测 周常胜(导师：姚卫红) - 《大连理工大学硕士论文》 - 2014-05-08 | 2.7% (73) 是否引证：否 |
| 4 | 基于特征向量的最小二乘支持向量机PM2.5浓度预测模型 李龙;马磊;贺建峰;邵党国;易三莉;相艳;刘立芳; - 《计算机应用》 - 2014-08-10 | 2.6% (70) 是否引证：否 |
| 5 | 量子混合蛙跳算法在过程神经网络优化中的应用 张强;许少华;刘丽杰; - 《信号处理》 - 2013-08-25 | 2.5% (68) 是否引证：否 |
| 6 | 分式过程神经网络在网络流量预测中的应用 张强;许少华;李盼池; - 《计算机工程与应用》 - 2012-12-11 | 2.5% (68) 是否引证：否 |
| 7 | 基于改进小波神经网络模型的交通流预测研究 陶丽(导师：刘升) - 《上海工程技术大学硕士论文》 - 2015-12-01 | 2.3% (63) 是否引证：否 |
| 8 | 卡尔曼滤波在捷联惯组误差系数预测中的应用 张焕鑫;李学锋; - 《计算机仿真》 - 2013-05-15 | 2.2% (58) 是否引证：否 |
| 9 | 善言为贤:货币政策前瞻性指引的中国实践 张成思;计兴辰; - 《国际金融研究》 - 2017-12-12 | 1.2% (31) 是否引证：否 |
| 10 | 基于快速离散曲波变换的TH-1卫星影像融合方法 朱倩;李霞; - 《中国科学技术大学学报》 - 2013-08-15 | 1.2% (31) 是否引证：否 |
| 原文内容 红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容 | | |

通过上述的研究分析，当在有离群点引入 ξ 之后，（4-10）可进一步变化得出如下公式：

对（4-23）做对偶变换：

得出拉格朗日方程，并给出分类决策函数：

4.2.4核函数

线性问题利用超平面分类可以很好解决，对非线性的情况则行不通。

如图4-4所示，将坐标轴中横轴（X轴）a、b之间即红色部分视为一个正类别，a和b两端的黑色部分视作一个负类。那么可以发现通过线性方法无法将两类分开，但是在（4-4）图中的蓝色曲线可以将两类别完美分开。

很明显，用线性的方法函数是无法对已有的数据样本进行区分的，所以引入核函数的方法。对核函数的方法思想前文也曾提及就是通过一种规则变换映射使得低维空间样本数据反映到高维，通过在高维的特征空间中利用线性方法进行区分。因此，在使用SVM或者SVR处理非线性情况，选择一个合适的核函数是解决问题的关键。经常使用的核函数有以下四类：

4.3基于支持向量回归预测模型研究

4.3.1模型参数分析

使用SVR模型算法，对核函数、特征空间以及非线性变换之间有相应的对应关系。对核函数的选择以及对核函数中参数的选法对非线性变换也会有影响，从而影响到算法复杂度。对如何选择核函数上当前并没有很好地统一标准，在总结前人大量的实践中本文采用的是高斯径向基核，它有以下特点：

（1）高斯核的适应范围很广，若选择了合适的参数，就可以有较宽的收敛域并且会得到很优良的性能以及很好的学习能力，适用于任一分布的数据样本，原始数据即可使用非线性反映到高维空间中。

(2) 核函数中参数的个数决定了算法模型选择的复杂度, 高斯核与多项式核或是多层感知器核比较, 它的参数量最少, 使用更为方便简单。

经过上述分析对比, 本文实验在搭建SVR算法模型使用的核函数即为高斯径向基核。则对SVR算法模型预测短时交通, 影响其预测效果的关键参数为: 惩罚参数 C 、不敏感损失系数 ϵ 及核参数 σ 。

a. 核参数 σ 对算法模型的影响

核参数 σ 值的大小对SVR性能的影响很大, 因为对支持向量的相关程度和训练数据的分布特性都是被核参数 σ 决定。若 σ 的值偏大则表明支持向量间的相关性越强, 模型推广能力会由于 σ 的增加反而会减弱, 则预测精度得不到保证; 若 σ 的值偏小则支持向量间的相关性会很弱, 模型的复杂性加强则推广能力无法得到保障。

b. 惩罚参数 C 对算法模型的影响

惩罚参数 C 是主要来调控置信区间范围和学习机的经验风险在数据子空间的比例大小, 希望使得算法结构能达到结构风险最小和很好的推广能力。算法模型的健壮性、复杂性以及在对那些在 ϵ 管道域外的样本点数据的惩罚程度的大小都是由 C 决定。因此, 在已有的数据空间中, C 值若太大, 则说明此时算法的推广性就会很差因为算法模型相对于数据的拟合度偏高, 出现了“过学习”现象; C 值若太小说明对惩罚经验误差轻, 训练数据的误差会变大, 这也就造成“欠学习”现象。对任一数据空间总会存在一个最优的惩罚参数 C 可以使得SVR具有很好地推广性能, 但若 C 越过了某范围, 则对SVR算法模型的推广力和降低经验风险方面失去意义。

c. 不敏感损失系数 ϵ

在算法中, 样本数据中存在的不敏感区域的范围、支持向量个数甚至模型推广性能都会受到不敏感损失系数 ϵ 的影响。如若 ϵ 取值过于小, 必然会使SVR模型复杂度增加, 预测效果会得到提高然而求解算法模型的时间会增加很多, 支持向量的数目也会增加, 那么就会出现“过拟合”, 此时模型在推广能力上会大打折扣; 相反如果 ϵ 值过大, 导致“欠拟合”, 模型的推广能力也会大打折扣。

通过对上述参数的研究分析可知, 想要SVR算法模型有很好的预测效果就必须合理选择 C 、 σ 、 ϵ 这三个参数值。

3.3.2 网格法选取模型参数

网格法的大体思想是通过一定的规则把算法模型的参数取值范围分割成若干个小区域, 接着通过计算出参数取值的所有可能组合并且计算出相应的目标误差, 然后进行一番比较选择在取值范围中目标值最小时对应的参数组合。这种选取参数的方法在理论基础上使得得出的解为范围中的一个全局最优解, 可以较少发生重大的误差。

对SVR算法使用高斯核的并通过网格法调参的步骤如下:

(1) 将惩罚参数 C 、核参数 σ 及不敏感系数 ϵ 分别按可能的取值范围通过固定的搜索补偿步长使用网格法进行划分, 并且他们的取值范围应当在2的指数空间上这样方便离散化搜索。

(2) 在通过(1)的基础上选出所有可能的参数之间的组合通过交叉验证的方法求出均方误差并比较, 选出均方误差最小的参数组。若在后续的搜索时发现有一组的均方误差与已经得出的最小误差相近并且其 C 的值还要更小时, 则原来的参数组合进行更新。

4.3.3 实验评价指标

为了更好地量化分析模型的预测性能本文中主要采用平均绝对误差 (Mean Absolute Error, MAE)、均方误差 (Mean Square Error, MSE)、均等系数 (Equal Coefficient)。

其中, $Y_p(t)$ 为 t 时刻模型预测值, N 为预测时段长度, $Y_r(t)$ 为 t 时刻交通流实际测量值。MSE反应误差分布情况, 值越小, 说明预测模型描述实验数据具有更好的精确度, 预测效果越好。EC反映预测值和实际测量值之间的拟合程度, 值越大越接近于1, 表示预测效果越好。

4.3.4 仿真实验

本文选择交通流的最常用参数流量作为预测对象。数据来源是前文中已经进行过预处理的预测样本。样本从2016年6月30日到2016年7月1日, 数据检测周期为1min, 预测间隔为10min, 选取其中08:00—22:00期间的数据每天共有79组数据分别进行训练和预测。结合使用的支持向量回归的短时交通流预测模型进行预测, 预测结果与实际记录值得结果对比图如下:

从预测结果结果对比图和表中数据来看, 采用网格搜索的方式可以达到一定的预测效果。网格算法以固定步长对参数空间的所有组合进行穷举实验, 对于参数较少的情况该方法取得了很好的效果, 但是在实际应用中该算法效率不高。因此本文将在下面对SVR模型参数的选择采取其他方法进行改进。

3.4 本章小结

本章首先介绍了支持向量机的基础理论和机器学习的相关知识, 其中包括了经验风险最小化, 以及结构风险最小化的理论。

在此基础上, 分析了支持向量机和回归机的原理, 建立 SVR 的模型, 并将预测结果与实测值进行对比, 来证明支持向量回归模型针对短时交通流的预测是一种可行的、有效的预测模型。为了进一步提高模型预测效果可以接近实际应用, 本文将在SVR基础上采取更适合在交通流预测中的参数选择方式进一步进行试验。

1. 其中， $Y_p(t)$ 为t时刻模型预测值，N为预测时段长度， $Y_r(t)$ 为t时刻交通流实际测量值。MSE反应误差分布情况，值越小，说明预测模型描述实验数据具有更好的精确度，预测效果越好。EC反映预测值和实际测量值之间的拟合程度，值越大越接近于1，表示预测效果越好。

4.3.4仿真实验

本文选择交通流的最常用参数流量作为预测对象。数据来源是前文中已经进行过预处理的预测样本。

2. 1min，预测间隔为10min，选取其中08:00—22:00期间的数据每天共有79组数据分别进行训练和预测。结合使用的支持向量回归的短时交通流预测模型
3. 达到一定的预测效果。网格算法以固定步长对参数空间的所有组合进行穷举实验，对于参数较少的情况该方法取得了很好的效果，但是在实际应用中该算法效率不高。

4. 3.4本章小结

本章首先介绍了支持向量机的基础理论和机器学习的相关知识，其中包括了经验风险最小化，以及结构风险最小化的理论。

在此基础上，分析了

5. 模型，并将预测结果与实测值进行对比，来证明支持向量回归模型针对短时交通流的预测是一种可行的、有效的预测模型。

说明：1.仅可用于检测期刊编辑部来稿，不得用于其他用途。

2.总文字复制比：被检测论文总重合字数在总字数中所占的比例。

3.去除引用文献复制比：去除系统识别为引用的文献后，计算出来的重合字数在总字数中所占的比例。

4.去除本人已发表文献复制比：去除作者本人已发表文献后，计算出来的重合字数在总字数中所占的比例。

5.指标是由系统根据《学术期刊论文不端行为的界定标准》自动生成的。

6.红色文字表示文字复制部分;绿色文字表示引用部分。

7.本报告单仅对您所选择比对资源范围内检测结果负责。

8.Email：amlc@cnki.net

 <http://e.weibo.com/u/3194559873>

 http://t.qq.com/CNKI_kycx