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。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。研究对象的选择上，由于高速公路的路网结果相对其他道路比较简单，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性能。

**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](#_Toc452843791)

[ABSTRACT II](#_Toc452843792)

[目录 III](#_Toc452843793)

[第1章 绪论 1](#_Toc452843794)

[1.1研究背景 1](#_Toc452843795)

[1.2国内外研究现状 1](#_Toc452843796)

[1.3本文工作 3](#_Toc452843797)

[第2章 基于DEM的数字流域特征及提取 4](#_Toc452843798)

[2.1 TesorFlow介绍 4](#_Toc452843799)

[2.2 交通状态介绍 4](#_Toc452843800)

[2.3提取样本 8](#_Toc452843803)

[2.3.1 数据准备 8](#_Toc452843804)

[2.3.2 工具介绍 8](#_Toc452843805)

[2.3.3 提取步骤 8](#_Toc452843806)

[2.3.4 构建样本 8](#_Toc452843806)

[2.4 小结 17](#_Toc452843807)

[第3章 流域相似性分析的相关算法 19](#_Toc452843808)

[3.1 主成份分析 19](#_Toc452843809)

[3.2 K-Means算法 20](#_Toc452843810)

[3.3 层次聚类算法 20](#_Toc452843811)

[3.4 相似流域分析实例 21](#_Toc452843812)

[3.4.1 数据准备 21](#_Toc452843813)

[3.4.2 数据分析 22](#_Toc452843814)

[3.5 小结 27](#_Toc452843815)

[第4章 基于DEM的相似流域查找系统的设计 28](#_Toc452843816)

[4.1 需求分析 28](#_Toc452843817)

[4.2 总体设计 28](#_Toc452843818)

[4.2.1 系统结构设计 28](#_Toc452843819)

[4.4.2 系统流程设计 30](#_Toc452843820)

[4.3 系统功能设计 30](#_Toc452843821)

[4.4数据库设计 31](#_Toc452843822)

[4.4.1河流相似性分析基本信息表 31](#_Toc452843823)

[4.4.2河流相似性性分析属性字典表 33](#_Toc452843824)

[4.5详细设计 33](#_Toc452843825)

[4.5.1 单流域搜索 33](#_Toc452843826)

[4.5.2 多流域分析 34](#_Toc452843827)

[4.5.3 DEM参数文件上传 35](#_Toc452843828)

[4.5.4 查询河流信息 36](#_Toc452843829)

[4.5.5 更新河流信息 36](#_Toc452843830)

[4.5.6 新增河流信息 36](#_Toc452843831)

[4.5.7 界面设计 37](#_Toc452843832)

[4.6 小结 38](#_Toc452843833)

[第5章 基于DEM的相似流域查找系统的实现 39](#_Toc452843834)

[5.1 系统开发环境 39](#_Toc452843835)

[5.2 系统实现 39](#_Toc452843836)

[5.2.1 单流域搜索功能实现 39](#_Toc452843837)

[5.2.2 多流域分析功能实现 41](#_Toc452843838)

[5.2.3 上传DEM参数文件功能实现 42](#_Toc452843839)

[5.2.4 河流信息管理功能实现 46](#_Toc452843840)

[5.2.5 系统界面实现 47](#_Toc452843841)

[5.3 系统展示 48](#_Toc452843842)

[5.3.1 系统界面展示 48](#_Toc452843843)

[5.3.2 单流域搜索功能 49](#_Toc452843844)

[5.3.3 多流域分析功能 50](#_Toc452843845)

[5.3.4 河流信息查询功能 51](#_Toc452843846)

[5.3.5 上传DEM参数文件功能 52](#_Toc452843847)

[5.3.7 新增河流信息功能 52](#_Toc452843848)

[5.3.8 编辑河流信息功能 53](#_Toc452843849)

[5.4 小结 53](#_Toc452843850)

[第6章 总结与展望 54](#_Toc452843851)

[致 谢 55](#_Toc452843852)

[参考文献 56](#_Toc452843853)

[附录1：国外经典论文——原文 58](#_Toc452843854)

[附录2：国外经典论文——翻译 69](#_Toc452843855)

**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。

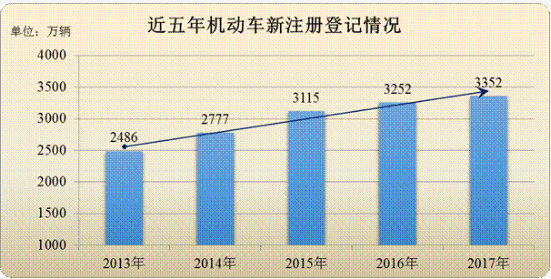


图1 机动车新用户

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。

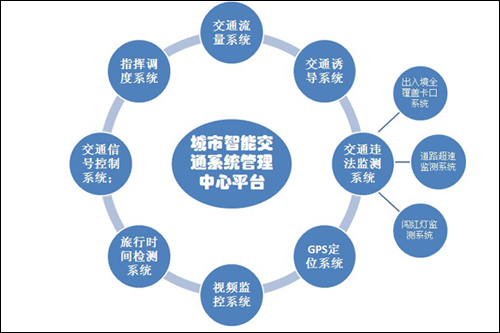


图2 智能交通系统概率图

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：

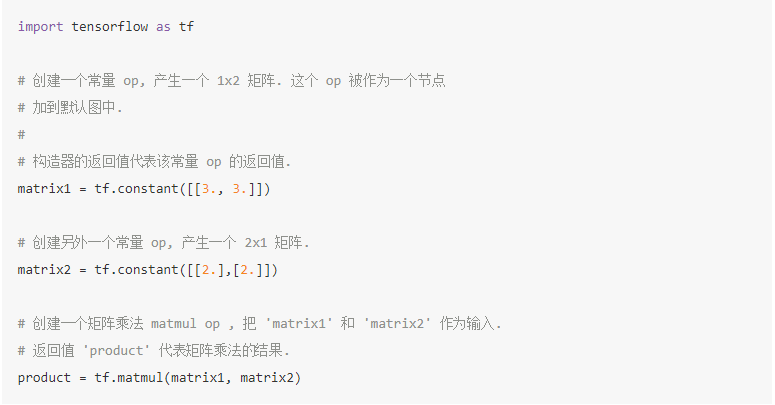


图3 tf中构建计算图的代码

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：

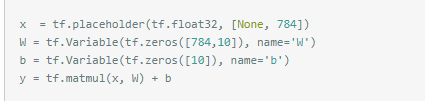


图5 tf中构建计算图

执行：

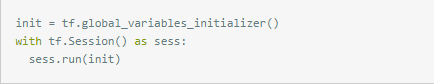


图6 tf中执行计算图

计算图：

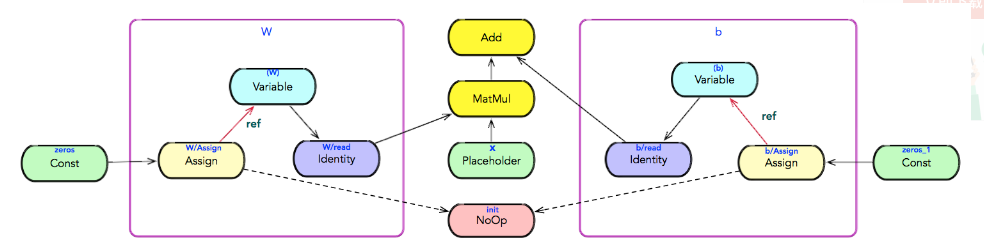


图7 构建的计算图模型

**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：

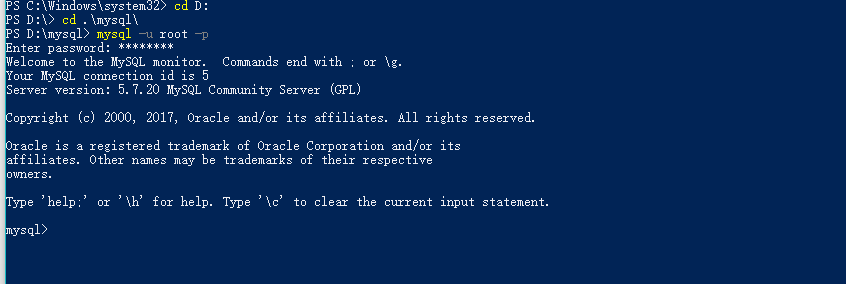


图8 PowerShell中打开MySQL

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：

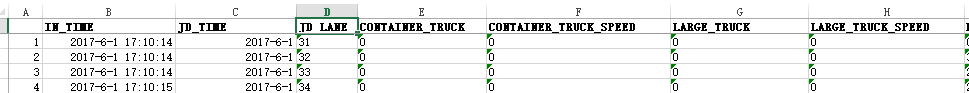


图9 原始的Excel数据（部分）

将Excel中的数据删除没用的列，再另存为.csv格式

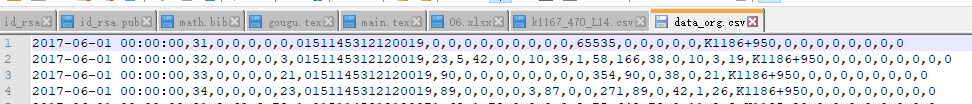


图10 删除几个没有列形成的.csv文件

在数据库中建相应的表：

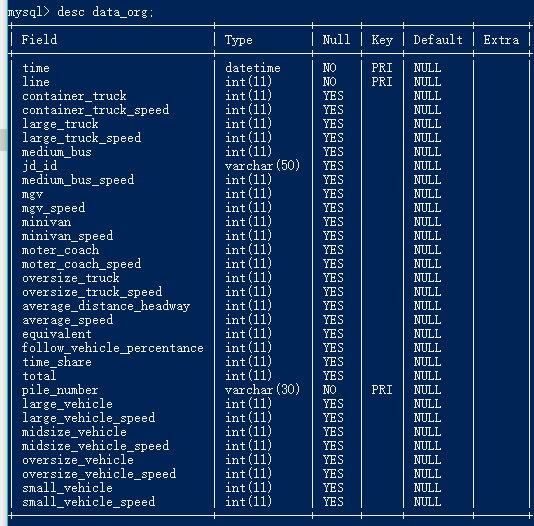


图11 数据库中用于导入的表的列属性

导入外部文件:

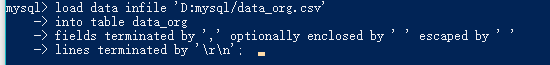


图12 导入代码文件的指令

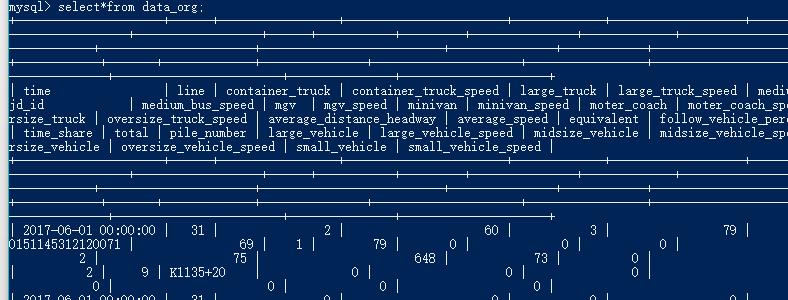


图13 导入之后的结果

再进行简单的SQL语句选择自己需要的数据，导出。

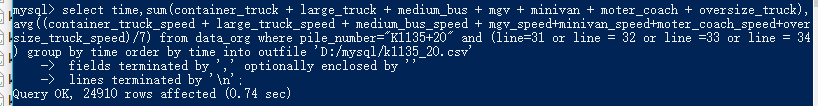


图14 查询车流量和车速

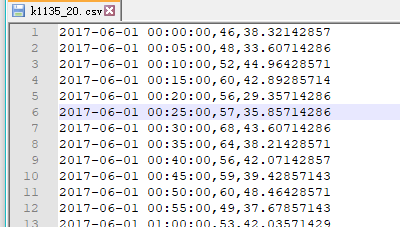


图15 导出的结果

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：

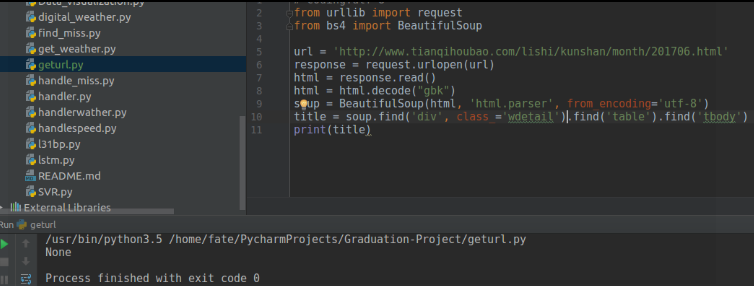


图19 静态提取天气

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：

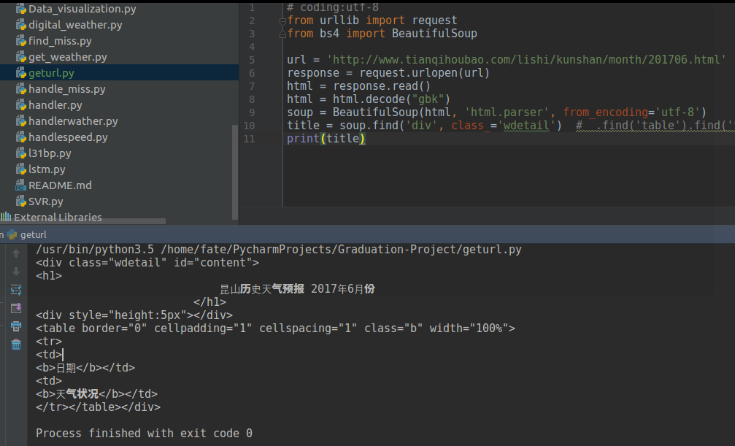


图20 提取div中内容

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：

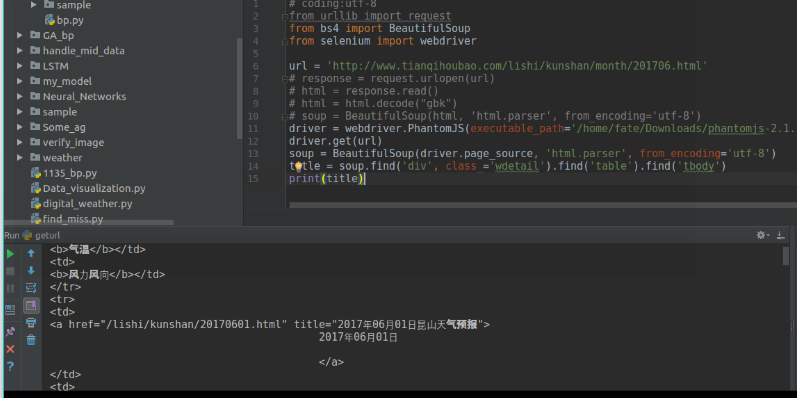


图21 动态提取天气

再把提取好的数据进行相应处理，如下显示：

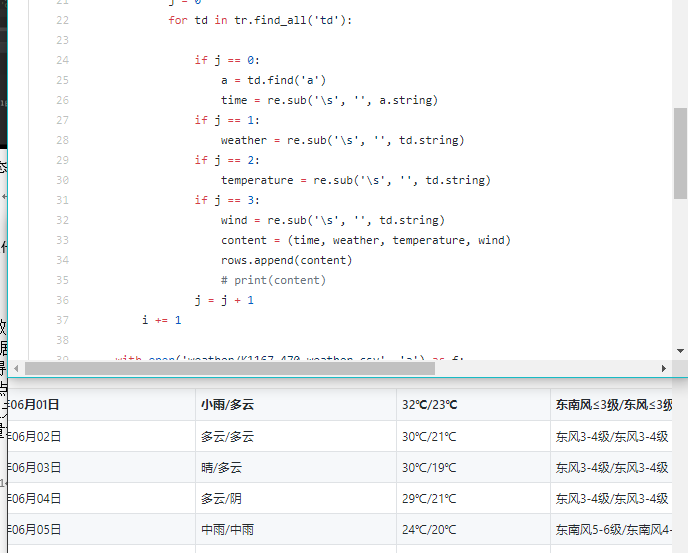


图22 处理代码和结果

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：

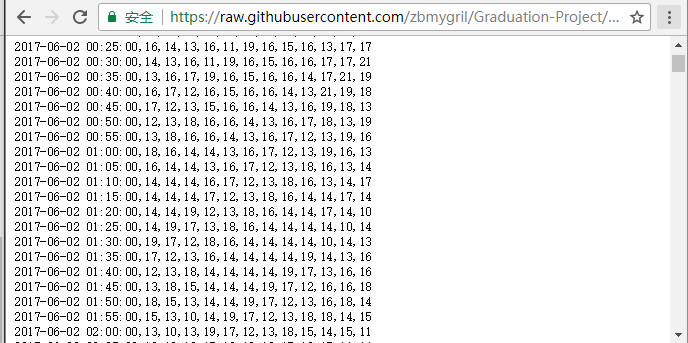


图23 交通流数据



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 。

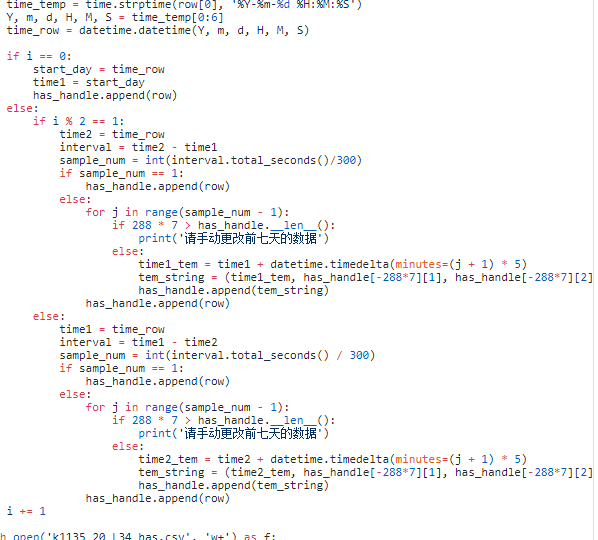


图25 处理丢失样本的代码



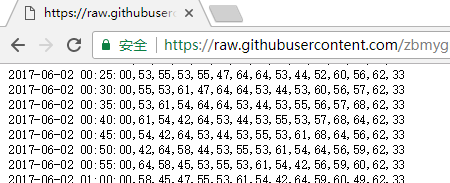
图26 合并交通流量和天气

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：

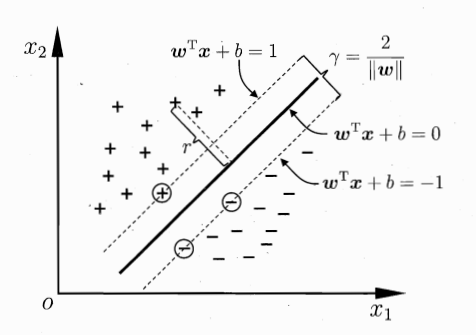


图28 SVM（截选周志华的机器学习）

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：

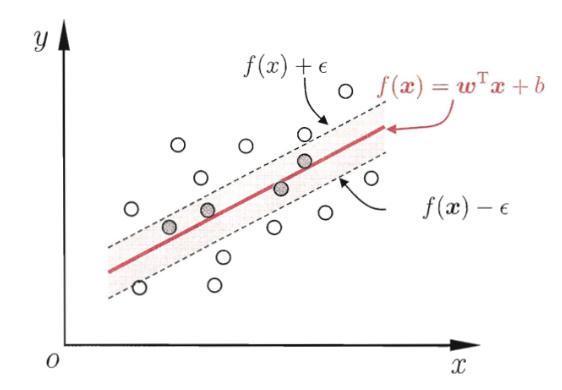


图29 SVR（知乎用户：石晓文的学习日记）

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、：

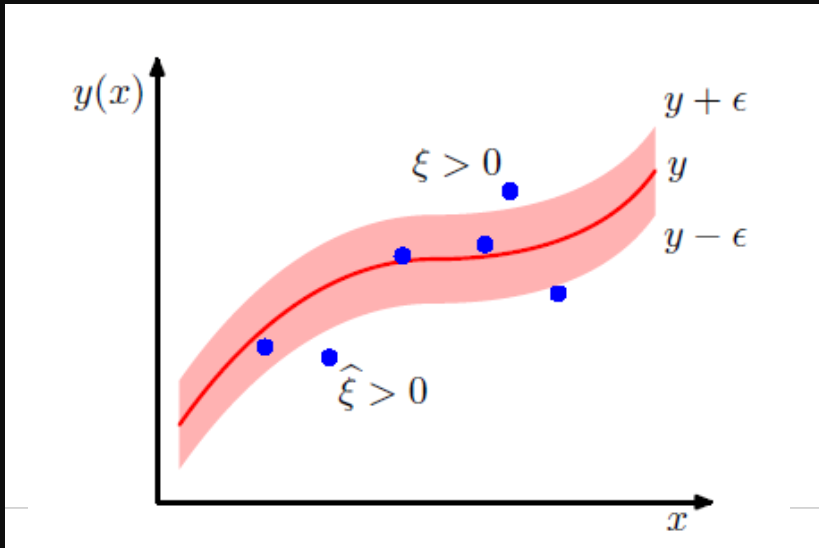


图30 引入松弛变量后的SVR（知乎用户：石晓文的学习日记）

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。

**3.2 K-Means算法**

K-Means方法是一个基于原型的划分聚类技术，它根据聚类的重心将样本划分成用户指定数量的类别。它将一个数据对象的集合划分成互不重叠的子集（集群），每个数据对象是一个确切的子集。在K-Means方法中，每个集群与质心（中心点）相关联，并且每个点会被分配到距离最近的质心的集群。其中，聚类的个数必须指定，记为K。它是一个高效的贪婪算法，一个代表性的聚类算法。

K-Means算法基本步骤如下：

（1）定义初始质心，其个数为聚类数量K。可以随机分配所有组的质心也可以自己定义K个不同的值作为实体的质心。

（2）将每一个点分配给最接近的质心所在的类，为了找出最相似的质心，该算法必须计算所有样本到每个质心间的距离。

（3）重新计算质心的值。

（4）重复第二步和第三步，直至集群不再发生变化。

可以看出，初始质心的选择对后期聚类的精度影响很大，为减轻这种影响，当样本数量较少时，可以先使用其他方法求出特定的K个质心的值；当样本数量非常多时，随机选择了最初的质心后，后续质心的选择尽量远离初始质心的位置。

**3.3 层次聚类算法**

层次聚类将数据构造成树型结构（通常是二进制）。叶子是单独的数据项，而根是一个包含所有数据的单簇。树叶和根中间是包含数据子集的中间簇。层次聚类的主要思路是使“集群的集群”向上构建一棵树。 阶层式汇聚分群法（HAC）从底部开始，每个成员作为一个单簇，合并组合到一起，最后全部成员组成根，或者某个定义终结的条件被满足。分裂聚类是从一个大集团的所有数据开始，不断下分，直到每个成员作为一个单簇存在，或者满足终止条件。

假如有N个样本，层次聚类的基本过程如下：

（1）将每一个样本归为一个单独的类；

（2）成对计算所有类的距离，通过距离度量方法求取，如门式距离、欧式距离、马氏距离等；

（3）依据所得距离值构造距离矩阵；

（4）寻找距离最短的两个类；

（5）从矩阵中删除这两个类，并将它们合并成一个类；

（6）计算新合并的类到其他所有类的距离，并更新矩阵；

（7）重复上述步骤，直到距离矩阵被减少到一个单一的元素，或者达到了自定义的终止条件。

聚类分析算法的关键是怎样量化两个集群间的异同。层次聚类中类的合并方法有以下几种：

1. 最短距离法：又称单锁链准则，从所有测得的两个类之间的距离中选择相距最短的两个类进行合并。
2. 最长距离法：也称完全连锁聚类分析，和最短距离法相反，它是根据两个类间的最长距离进行类的选择合并。
3. 中间距离法：将两个类中的对象的距离求中间值，基于此中值进行合并。
4. 类平均法：将两个类中的对象的距离全部放在一起求平均值，基于此平均值进行合并。
5. 重心法：用比较所有类质心的差异来代替计算类间的距离作为合并标准。

层次聚类具有助于数据显示、分类细致的优点，但它没有提供能在早期阶段中“错误”对象的重定位，结果需要仔细核对以确保它符合条理。而且使用不同的距离公式测量簇集间距离时可能会产生不同的结果，建议进行多次实验并比较结果。

**3.4 相似流域分析实例**

本节以江西26个小流域为例，利用R语言进行流域相似性分析。

**3.4.1 数据准备**

原始数据中包括安和、大斜、东村、渡头、鹅公湾、高安、洪家园、居龙滩、李家渡、莲花、良田、麟潭、梅港、南关口、三都、上高、深渡、田头、万家埠、峡山、烟头、杨树坪、窑棚、宜丰、永修和樟树这26个流域出口站点的流域面积、流域长度、流域平均坡度、形态因子、伸长比、河网密度、河道维持常数、平均河链长、河链平均汇水面积、河网总长度、近似常数k、河流频度、河链频度、常流性主河道长度、常流性主河道比降、最大流路距离、流域高程曲线面积、面积坡度这18个基于DEM提取的流域特征值和1月平均降雨量、2月平均降雨量、3月平均降雨量、4月平均降雨量、5月平均降雨量、6月平均降雨量、7月平均降雨量、8月平均降雨量、9月平均降雨量、10月平均降雨量、11月平均降雨量、12月平均降雨量、多年平均降雨量这13个水文参数指标。

**3.4.2 数据分析**

将原始数据分别使用18个基于DEM提取的流域特征值作为特征因素（记为R1）进行分析。

1. 调用R语言的stats包对特征因素进行主成份分分析。

首先用fa.parallel()分析主成分个数,结果如下：R1经平行分析后建议取三个主成份。

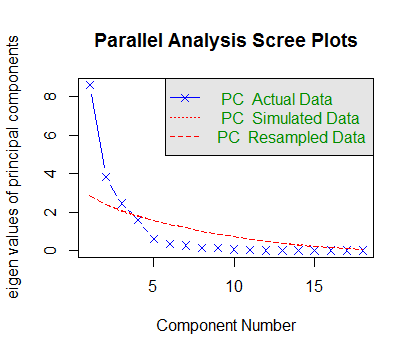


图3.1 R1主成份个数分析图

再使用principal函数对特征参数值进行降维,将R1的18个指标变换成3个独立的主成份因子。

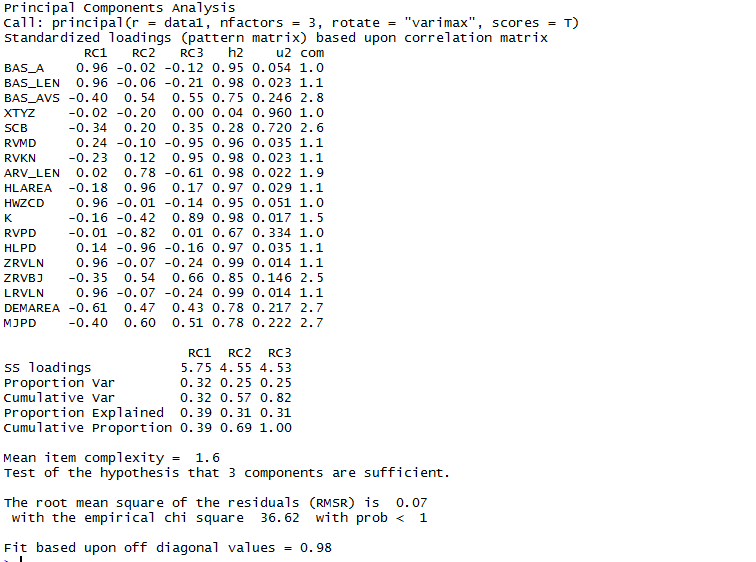


图3.3 R1主成份分析图

据图3.3所示，此处输入的是原始数据，并指定获取最大方差旋转的主成份，由于PCA值对相关系数矩阵进行分析，原始数据会自动被转换成相关系数矩阵，RC1-RC3栏包含了成分载荷（component loadings），表示因子与主成份间的相关系数。SS loading包含了主成份相关联的特征值，Proportin Var表示的是每个主成分对整个数据集的解释程度，第一主成分PC1解释了32%的方差，第二主成份PC2解释了25%的方差，第三主成份PC3解释了25%的方差。

1. 调用R语言中的层次聚类方法对经过主成份分析的数据进行分析。

表3.1 R1的三个主成份及其特征值

|  |  |  |  |
| --- | --- | --- | --- |
|  | 第1主成份 | 第2主成份 | 第3主成份 |
| 安和 | -0.81661 | 1.429826 | -0.49474 |
| 大斜 | -0.28889 | 0.621471 | 0.587337 |
| 东村 | 0.084263 | -0.79265 | -0.05008 |
| 渡头 | -0.96237 | -1.87727 | 0.431938 |
| 鹅公湾 | -0.67875 | -1.03855 | 0.212771 |
| 高安 | 0.005225 | -0.10227 | -0.71684 |
| 洪家园 | -0.06068 | 0.414479 | -1.08463 |
| 居龙滩 | -1.0824 | 0.062121 | -1.67467 |
| 李家渡 | 0.038787 | -0.44837 | -1.07827 |
| 莲花 | -0.29905 | 0.133916 | 0.325388 |
| 良田 | -0.68506 | -1.54256 | 0.927939 |
| 麟潭 | -0.43916 | 1.333999 | 1.839812 |
| 梅港 | -0.91306 | -0.10468 | -1.4895 |
| 南关口 | -0.65189 | -1.24403 | -0.53059 |
| 三都 | 0.311867 | 0.792925 | -0.65493 |
| 上高 | 0.795871 | -1.06032 | -0.4073 |
| 深渡 | 0.499737 | -0.22671 | 0.902744 |
| 田头 | -0.31695 | 0.920094 | 2.644591 |
| 万家埠 | 1.917708 | -0.03666 | 0.397815 |
| 峡山 | -1.04333 | 2.725364 | -0.68368 |
| 烟头 | -0.19657 | 0.364906 | 0.82093 |
| 杨树坪 | 0.264899 | -0.06064 | 0.253608 |
| 窑棚 | -0.25068 | -0.906 | 0.826998 |
| 宜丰 | -0.28754 | 0.050893 | -0.2453 |
| 永修 | 1.897787 | 0.316895 | -1.07454 |
| 樟树 | 3.156856 | 0.273841 | 0.013193 |

先计算数据间的距离，调用dist（）函数，method=“euclidean”，即欧几里德距离作为公式计算。然后，调用hclust()函数，分别使用其中的 “complete”、“single”、“average”三种距离方法对处理过的数据进行分析，画出聚类分析图。

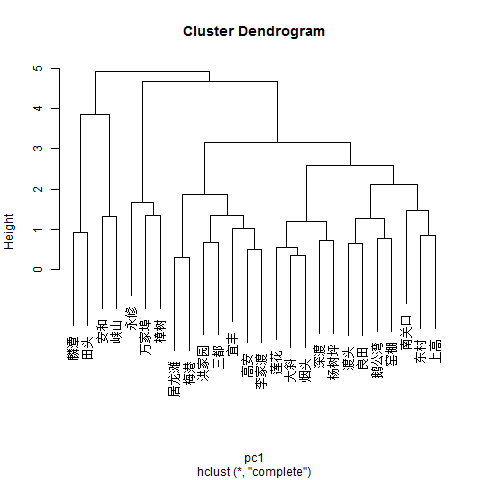


图3.4 最长距离法聚类分析

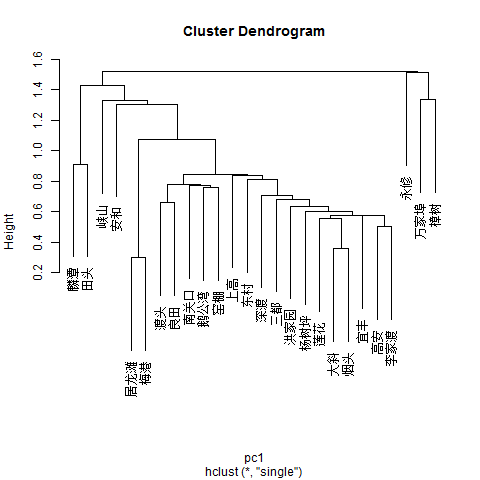


图3.5 最短距离法聚类分析

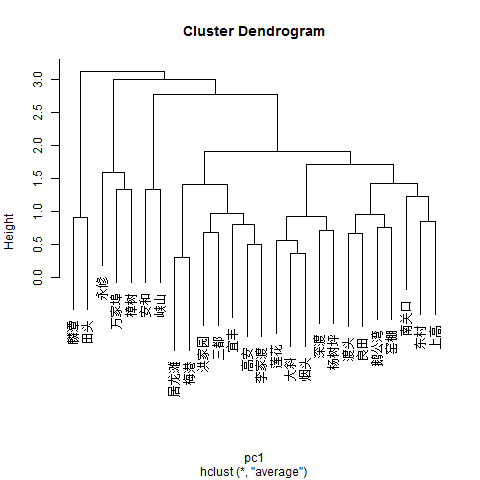


图3.6 平均距离法聚类分析

根据图3.4-3.6三张聚类结果图，我们可以发现万家埠、永修和樟树三个站在三种算法中都表现出较高的相似度，而南关口站在平均距离算法和最长距离算法中，和上高、东村两个站结果较为相似，但在最短距离算法中，它和窑棚、鹅公湾两个站结果较为相似。这表明层次聚类算法在使用不同的距离公式测量簇集间距离时可能会产生不同的结果，应多做几次实验进行比较。

1. 调用R语言Knn包中的K-Means方法对经过主成份分析的数据进行分析。

调用kmeans()函数，分别使用其中的 “MacQueen”、“Forgy”、“Lloyd”三种距离方法对处理过的数据进行分析，研究发现其结果中的规律。

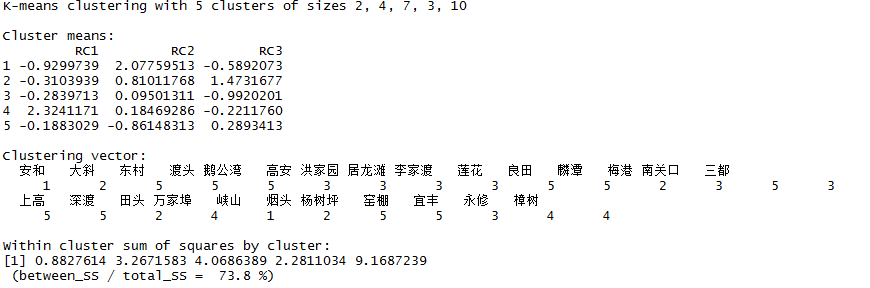


图3.7 Forgy方法

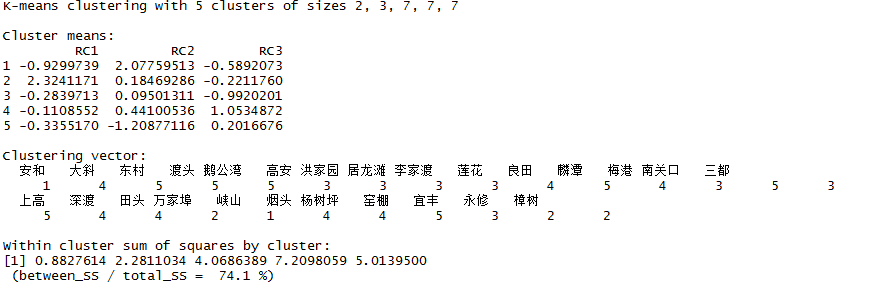


图3.8 MacQueen方法

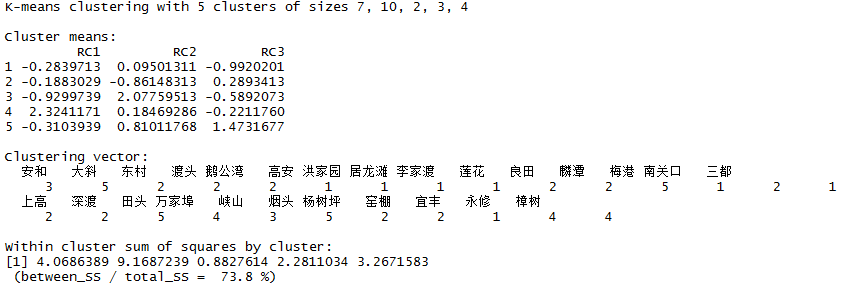


图3.9 Lloyd方法

Clustering vector表示每个样本所属的聚类，在一个聚类中的样本被认为是相似的，被划分成了一个类型。Within cluster sum of squares by clusters代表聚类内部的距离平方和。Between\_SS/total\_SS表示不同聚类间距离平方和及整体距离平方和的比值，比值越高，算法精确度越高。将上述结果进行比较发现，Forgy方法和Lloyd方法对26个站点的分类结果是相同的，但它们的聚类内部的距离平方和不相同，三个方法的精确值相差不大。

**3.5 小结**

本章主要对流域相似性分析的主成份分析步骤和两种基本算法进行了理论上的说明，并用江西省的26个出口站点信息作为数据进行了实例分析，对算法有了进一步的了解。为下文实现相似流域查找系统提供了技术支持。

**第4章 基于DEM的相似流域查找系统的设计**

我国幅员辽阔、地貌复杂，河域支流繁多，导致流域多样化。近年来，我国大江大河在洪水预报、流域治理方面取得了不错的成果。但众多的中小流域因流域资料不全或测站缺乏导致没有流域资料，在防洪防灾方面收效胜微。相似流域的查询对流域资料叠加整合、水文资料补全有着重要意义。因此，从数字高程模型提取有效特征参数，建立健全基础流域信息系统，整合共享有效信息对于中小流域治理有深远影响。本文基于两种相似性查询方法，设计并实现了基于DEM的相似流域查找系统。

**4.1 需求分析**

本文使用的数据分为两部分，一部分是从数字高程模型中提取到数字流域特征参数文件，格式为txt；另一部分来源于国家水文数据，数据类型涉及各个水文站的站码、经纬度、站名、行政区划以及雨量信息。本系统——基于DEM的相似流域查找系统的目标是：根据以上数据，通过一系列预处理之后，借助层次聚类、K-Means等算法，通过分析各个流域的水文数据，挖掘其中流域之间的相似性，并以图片的形式形象的展示给水利工作者。这些相似性信息能够帮助水利工作者分析水情，帮助水利工程设计者针对将要开展的水利工程寻找有效的参照，帮助流域监察者更好的处置突发情况。也为无资料或缺资料地区提供了更多能够参考的信息，对中小流域预防各种自然灾害有着重要意义。将来希望能够形成一个全国各流域信息的展示和比较平台，它的实现将能够给各类水利相关人员在处理问题时，提供相似流域针对相似问题的处理过程，以此作为参考，帮助其解决问题。

依据系统构建规划，本系统的主要功能包括：

（1）实现对流域特征数据以及DEM特征信息文件的增加和更新。

（2）实现对数据库中的流域信息数据的查询和预处理。

（3）实现对预处理结果，能够以两种不同的相似性分析方法进行分析。

（4）实现将分析结果以图表和文字等多种形式展现给用户。

**4.2 总体设计**

**4.2.1 系统结构设计**

基于DEM的流域相似性查找系统需要从数据库中读取数据进行分析并将分析结果以某种形式反馈给用户，适合基于J2EE平台的MVC模式来实现。MVC模式把用户与系统的交互分成三个不同的角色，各自独立处理自己的任务：Model处理应用程序中与数据相关的部分，通常负责在数据库中读写数据；View处理数据显示相关的部分，通常直接从模型中取得它需要显示的状态与数据；controller处于View和Model的中间，负责接受用户的输入，将输入进行分析并反馈给模型。根据上述要求，可将基于DEM的相似流域查找系统结构分为三个层次进行设计，如图所示：

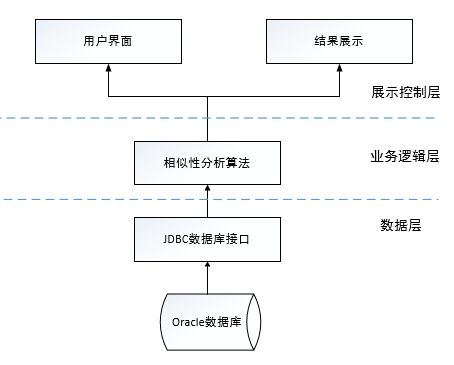


图4.1 系统层次结构图

1. 展示层: 其功能是接收用户提交的文件和输入的查询信息等前台输入，当后台处理分析后，处理结果会返回该层被展示，由Jsp和JavaScript组件实现。
2. 控制层：其功能是接受并处理从展示层收到的任务信息，并调用业务逻辑层的代码进行实现。由Servlet组件实现。
3. 业务逻辑层：根据Servlet发来的指令，从数据层的数据库中获取所需数据，进行分析，得到结果后，将其反馈给表示层进行展示，由相似性分析算法组件构成。
4. 数据层：根据业务层发送的请求读取数据到业务层中，由数据库和数据接口组成。

**4.4.2 系统流程设计**

系统运行是一个接收、处理、完成用户请求的循环过程。这期间，对于用户发布的不同类型的请求，系统会加载相应的模块去解决。系统整体流程设计如下图所示。

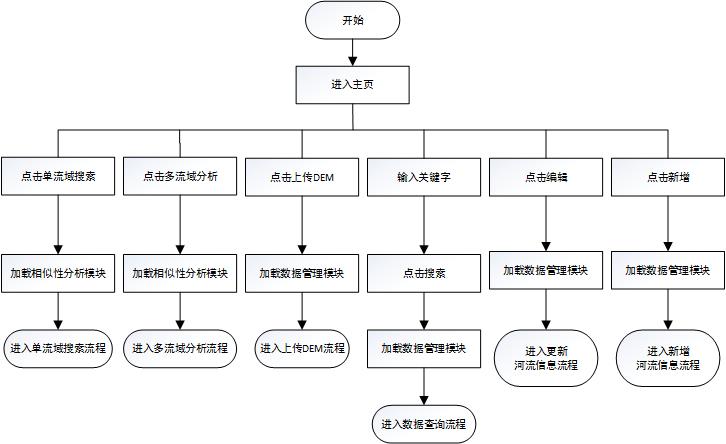


图4.2 系统流程图

**4.3 系统功能设计**

基于DEM的相似流域查找系统是一个深度挖掘DEM有效信息，查询分析河流水文信息的系统。本系统主要包括相似性分析和数据管理两个功能模块。其中相似性分析模块分为单流域搜索和多流域分析两种方式。而数据管理模块可以添加、更新和查询河流信息和上传DEM参数文件。基于DEM的相似流域查找系统功能结构如下图4.3所示。



图4.3 系统功能结构图

**4.4数据库设计**

基于DEM的相似流域查找系统数据来源于国家水文数据库和DEM高程数据。由于全国的水文数据量庞大，需要一个相对稳定的存储环境，而Oracle在安全性、稳定性方面优于其他数据库，且Oracle在数据导入和数据多种格式支持方面功能强大，所以选择Oracle 11g作为数据存储的工具。系统涉及到以下两张表：河流相似性分析基本信息表和河流相似性性分析属性字典表。

**4.4.1河流相似性分析基本信息表**

表标识为SIMILAR\_DATA，该表主要存储河流基本信息和流域特征信息。内容如下表所示：

表4.1 河流相似性分析基本信息表主要表结构

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 序号 | 字段名 | 标识符 | 类型及长度 | 有无空值 | 计量单位 | 主键 |
| 1 | 河流编码 | RVCODOE | nvarchar(12) | N |  | Y |
| 2 | 河流名称 | RVNAME | nvarchar(100) | N |  | Y |
| 3 | 行政区划 | ADDVCD | nvarchar(100) | N |  |  |
| 4 | DEM状态 | DEMFLAG | nvarchar(10) | N |  |  |
| 5 | 时间戳 | TIMESTAMP | nvarchar(20) | N |  |  |
| 6 | 类型 | TYPE | nvarchar(20) |  |  |  |
| 7 | 流域面积 | BAS\_A | number(8,2) |  | km^2 |  |
| 8 | 流域长度 | BAS\_LEN | number(9,4) |  | km |  |
| 9 | 流域平均坡度 | BAS\_AVS | number(15,8) |  | ° |  |
| 10 | 形态因子 | XTYZ | number(8,4) |  |  |  |
| 11 | 伸长比 | SCB | number(8,4) |  |  |  |
| 12 | 河网密度 | RVMD | number(8,4) |  | km/km^2 |  |
| 13 | 河道维持常数 | RVKN | number(8,4) |  | km^2/km |  |
| 14 | 平均河链长 | ARV\_LEN | number(8,4) |  | km |  |
| 15 | 河链平均汇水面积 | HLAREA | number(9,4) |  | km^2 |  |
| 16 | 河网总长度 | HWZCD | number(8,2) |  | km |  |
| 17 | 近似常数k | K | number(8,4) |  |  |  |
| 18 | 河流频度 | RVPD | number(8,4) |  | 1/km^2 |  |
| 19 | 河链频度 | HLPD | number(8,4) |  | 1/km^2 |  |
| 20 | 常流性主河道长度 | ZRVLN | number(8,2) |  | km |  |
| 21 | 常流性主河道比降 | ZRVBJ | number(8,4) |  | m/m |  |
| 22 | 最大流路距离 | LRVLN | number(8,2) |  |  |  |
| 23 | 流域高程曲线面积 | DEMAREA | number(8,4) |  |  |  |
| 24 | 面积坡度 | MJPD | number(8,4) |  | dm/km^2 |  |
| 25 | 1月平均降雨量 | JARAVP | number(8,2) |  |  |  |
| 26 | 2月平均降雨量 | FEBAVP | number(8,2) |  |  |  |
| 27 | 3月平均降雨量 | MARAVP | number(8,2) |  |  |  |
| 28 | 4月平均降雨量 | APRAVP | number(8,2) |  |  |  |
| 29 | 5月平均降雨量 | MAYAVP | number(8,2) |  |  |  |
| 30 | 6月平均降雨量 | JUNAVP | number(8,2) |  |  |  |
| 31 | 7月平均降雨量 | JULAVP | number(8,2) |  |  |  |
| 32 | 8月平均降雨量 | AUGAVP | number(8,2) |  |  |  |
| 33 | 9月平均降雨量 | SEPAVP | number(8,2) |  |  |  |
| 34 | 10月平均降雨量 | OCTAVP | number(8,2) |  |  |  |
| 35 | 11月平均降雨量 | NOVAVP | number(8,2) |  |  |  |
| 36 | 12月平均降雨量 | DECAVP | number(8,2) |  |  |  |
| 37 | 多年平均降雨量 | MYAVP | number(8,2) |  |  |  |

1）河流编码：按《国家水资源监控能力建设项目标准——信息分类及编码规定》4.1 节执行。

2）河流名称：河流编码对应河流的名称。

3）行政区划：是国家为了进行分级管理而实行的区域划分，按照《[宪法](http://baike.baidu.com/view/3575.htm)》规定中国行政区划分为省、县、乡三级，这里指的是河流所属的行政区划。

4）DEM状态：标明DEM数据是未上传状态还是已上传状态。

5）时间戳：此记录第一次录入数据库的时间或者数据更新时间，取系统日期时间。

6）类型：指流域的大小情况，如中型流域、小型流域。

7）—24）为DEM数据中提取的属性值。

25）—37）为水文相关属性值。

**4.4.2河流相似性性分析属性字典表**

表标识为SIMILAR\_INDEX，该表是一张字典表，主要对河流相似性分析基本信息表SIMILAR\_DATA中的字段信息进行说明，对SIMILAR\_DATA中属性在前台显示的顺序进行定义。内容如下表所示：

表4.2河流相似性性分析属性字典表主要表结构

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 序号 | 字段名 | 标识符 | 类型及长度 | 有无空值 | 计量单位 | 主键 |
| 1 | 属性标识符 | CODE | NVARCHAR2(10) | N |  | Y |
| 2 | 属性名称 | NAME | NVARCHAR2(10) |  |  |  |
| 3 | 数据单位信息 | UNIT | NVARCHAR2(10) |  |  |  |
| 4 | 类型 | TYPE | NVARCHAR2(10) |  |  |  |
| 5 | 前台显示顺序 | ORDERBY | NVARCHAR2(2) |  |  |  |

**4.5详细设计**

**4.5.1 单流域搜索**

单流域搜索应实现如下功能：

1. 页面展示数据库中存储的河流信息；
2. 选出一条河流作为基准河流，这条河流的选择范围应该是数据库表存储的全部河流信息，基准河流可从河流数据表中勾选；
3. 与基准河流比较的河流的数量不限；
4. 因不能在一页中完全显示数据库中所有河流信息，河流信息表应分页显示；
5. 当选择的待比较的河流数量较多，而用户想要更新比较范围时，一键重置比较范围，将之前所选河流清空；
6. 当用户选择错少数待比较河流时，可以移除错误河流；
7. 因相似性指标含有18个DEM要素和13个水文要素，相对较多，用户可以选择自由组合这些要素进行相似性分析；
8. 可以按照行政区划显示处于该行政区的河流信息；
9. 用户可以自己调整Kmeans的聚类数目，比较不同的聚类数目下分析结果的异同；
10. 分析结果以图形、表格、文字等形式在前台界面展示。

单流域搜索的流程如下图所示：



图4.4 单流域搜索流程图

**4.5.2 多流域分析**

多流域分析应实现如下功能：

1. 页面展示数据库中存储的河流信息；
2. 待分析的河流数量不限，河流的选择范围应是数据库表存储的全部河流信息，可从河流数据表中勾选；
3. 因不能在一页中完全显示数据库中所有河流信息，河流信息表应分页显示；
4. 当选择的待分析的河流数量较多，而用户想要更新全部范围时，一键重置，将之前所选河流清空；
5. 用户能够移除已选的河流；
6. 用户可以自由选择组合相似性分析要素进行分析；
7. 可以按照行政区划显示处于该行政区的河流信息；
8. 用户可以自己调整聚类数目，比较不同的聚类数目下分析结果的异同；
9. 分析结果以图形、表格、文字等形式在前台界面展示。

多流域分析流程如下图所示：



图4.5 多流域分析流程图

**4.5.3 DEM参数文件上传**

当一些河流缺少DEM参数信息时，用户可以按照系统提供的文件格式上传DEM参数文件。因按照本文第二章所述提取DEM流域特征方法得到的DEM参数文件中除去本文使用的18个DEM特征指标外，还有归一化面积、归一化高程等大量其它数据。用户上传DEM参数文件后，先保存在服务器上，系统需要对文件进行分析，将系统所需要的参数提取出来，存入数据库中。因上传文件的名字并不统一，为了方便将提取的流域的DEM参数与河流信息匹配，把文件的名字命名为河流代码。

DEM参数文件上传流程如下图所示：



图4.6 DEM参数文件上传流程图

**4.5.4 查询河流信息**

输入河流代码或者河流名字等关键词，能在河流信息表中查询到相对应的河流信息。

**4.5.5 更新河流信息**

界面上应该给出一个编辑按钮，当点击编辑按钮时会弹出更新信息的界面。选择想要更新的河流，点选编辑按钮，输入修改的河流信息后确认更新。

**4.5.6 新增河流信息**

界面上应该给出一个新增按钮，当点击按钮时会弹出新增数据的界面。新增数据的界面应该包括河流编码、河流名称、行政区划、DEM状态、时间戳、类型、流域面积、流域长度、流域平均坡度、形态因子、伸长比、河网密度、河道维持常数、平均河链长、河链平均汇水面积、河网总长度、近似常数k、河流频度、河链频度、常流性主河道长度、常流性主河道比降、最大流路距离、流域高程曲线面积、面积坡度、1-12月平均雨量和多年平均雨量这些属性信息的输入框。

**4.5.7 界面设计**

基于DEM的相似流域查找系统查询方法界面UI设计如图4.7所示，顶部为“系统logo区”，其下方是“相似性查询功能选择区”，左上部分是“行政区划列表”，左下部分则是“河流范围选择区”，右半部分主要是“数据显示及结果展示区”和“相似性要素选择区”。其中“相似性要素选择区”是一个下拉框。



图4.7 界面设计图

1. 系统logo区：基于DEM的相似流域查找系统的名字和标识图片；
2. 相似性查询功能选择区：选择单流域搜索功能和多流域分析功能；
3. 行政区划表：对全国的城市按省→市→县（区）的级别进行分级显示，点击该城市，数据展示页面就会显示该城市的河流信息；
4. 河流范围选择区：单流域搜索在这个区域确定待搜河流和搜索范围，多流域分析在这个区域确定河流分析范围，即需要分析的全部河流在这个区域确定；
5. 相似性要素选择区：选择需要的相似性要素，共31个可供选择，包括18个DEM特征参数和13个水文要素；
6. 数据显示及结果展示区：以表格的形式展示河流信息数据，在查询结果返回后以图形和表格的形式进行展现。

**4.6 小结**

本章主要对基于DEM的相似流域查找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、设计了数据库表，最后基于上述工作，对系统进行了详细设计。

**第5章 基于DEM的相似流域查找系统的实现**

**5.1 系统开发环境**

* 操作系统：Windows 7
* 集成开发环境：ideaIU-13.1.6
* JAVA版本：JDK 1.7，JAVA EE 1.7
* Web 应用服务器：Tomcat-7.0.68
* 浏览器：IE 11，Firefox 11.0，Chrome
* 数据库：ORACLE 11g

**5.2 系统实现**

**5.2.1 单流域搜索功能实现**

根据上一章对单流域搜索模块的设计，当用户从前台页面选择了基准河流和与其进行比较的河流范围，确定了需要的相似性分析要素，点击确定后，后台的DataQueryServlet会根据前台传过来的一个flag进行判断。

当该flag的值为"singlerv"时，从前台页面获取singleCode、singleListCodes、checkboxsCodes、k四个参数，它们分别代表了待搜流域、待搜范围、要素范围和聚类数目。然后根据这些参数信息，调用SimilarResult类中的getMultiRes( )方法和ClusterDao类中的getclusterData( )方法从数据库中取出聚类需要用的数据。控制台再调用R语言的Kmeans方法进行分析,并将结果返回前台页面。其中，使用R语言调用Kmeans聚类算法是本节实现的重点。

上述过程中使用到以下几个类，及它们之间的关系如下图所示：

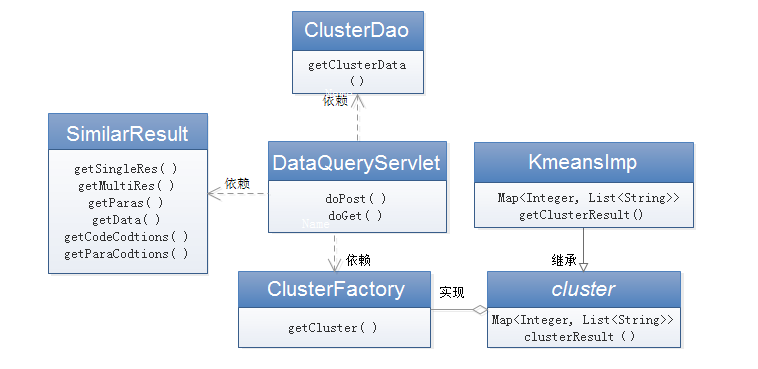


图5.1 单流域搜索uml类图

DataQueryServlet：控制台，对前台传来的flag判断类型后，调用各种类实现操作；

ClustDao：用于从数据库中获取分析需要的河流信息；

Cluster：调用R语言分析方法的接口；

ClusterFactory：实现Cluster的类；

KmeansImp：cluster的子类，调用R语言中Kmeans方法的具体代码。

本系统先连接Rserve服务器，向服务器发送需要分析的数据，调用写好的R语言脚本进行分析后，Rserve服务器将结果返回给控制台。JAVA调用Kmeans方法的脚本如下：

|  |
| --- |
| RserverConf rconf = new RserverConf("127.0.0.1", 6311, "conan", "conan", new Properties());  Rsession s = Rsession.newInstanceTry(System.out, rconf);  // 从文件加载R环境  String filePath = "E:/R/code/inputK.R";  s.eval("source(\"" + filePath + "\",encoding = 'UTF-8')"); |

K-means聚类过程中先要对相似性要素进行主成份分析，调用R语言psych包中的方法进行实现，代码如下：

|  |
| --- |
| #PCA降维  pc<-principal(basins, nfactors = k, score = T, rotate = "varimax")  means <- apply(basins,2,mean)  for (i in 1:ncol(basins)) {  if(i == 1){  basinsNew = data.frame(basins[,1] - means[1])  }else{  basinsNew = data.frame(basinsNew,basins[,i] - means[i])  }  names(basinsNew)[i] = names(basins)[i]  }  x<-as.matrix(basinsNew)  y<-pc$loadings  mainBasins <- x%\*% |

再进行K-means聚类分析：

|  |
| --- |
| fit\_kml <<- kmeans(mainBasins,center = k)  betweenss <<- fit\_kml$betweenss  totss <<- fit\_kml$totss #找出selectedCountry在countries中的位置  selectedRvindex=which(Basins$RVCodes == selectedrv)  #找出selectedCountry在cluster中的类别  cluster\_aim=fit\_kml$cluster[selectedRvindex]  #找出同属于cluster\_aim所有河流的index  similarityIndexs=which(fit\_kml$cluster == cluster\_aim)  #查找到的相似的河流  similarRvs <<- Basins[similarityIndexs,] |

因结果维度较高，当以图片的形态展示分析结果时，经MDS降维后，运用ggplot包中的方法画图并保存。

|  |
| --- |
| mds = cmdscale(dist(mainBasins,method = 'euclidean'),k=k,eig=T)  x = mds$points[,1] #MDS降维  y = mds$points[,2]  p <- ggplot(data.frame(x,y),aes(x,y,label=c(RVNames)))#画图 |

**5.2.2 多流域分析功能实现**

多流域分析功能的实现与单流域搜索的过程相似，都是通过调用R语言中的分析方法进行实现，其实现过程中用到的类和它们的关系如下图所示：

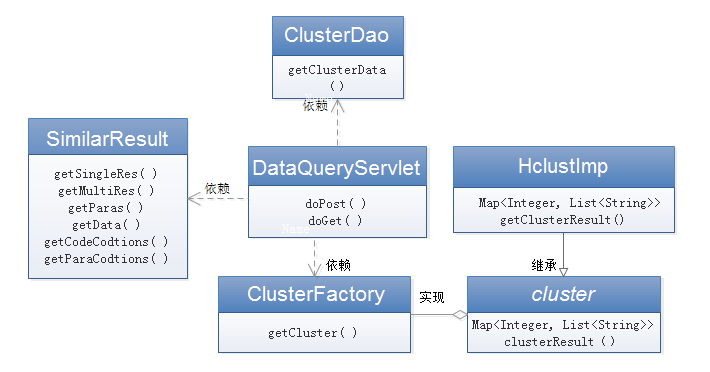


图5.2 多流域分析uml类图

DataQueryServlet进行判断并获取前台传输参数的代码如下：

|  |
| --- |
| if (flag.equals("multirv")) {  String multiListCodes = request.getParameter("multiListCodes");  String checkboxsCodes = request.getParameter("checkboxsCodes");  String kStr = request.getParameter("k");  String method = request.getParameter("method"); |

这里从前台得到的参数分别是multiListCodes多流域搜索的范围、checkboxCodes相似性分析参数、k聚类数目和聚类方法。然后根据这些参数信息，调用SimilarResult类中的getMultiRes( )方法和ClusterDao类中的getclusterData( )方法从数据库中取出聚类需要用的数据。控制台再调用R语言的方法进行分析,并将结果返回前台页面。

层次聚类算法的实现代码如下：

|  |
| --- |
| compoents <- fa.parallel(basins, fa = "pc", n.iter = 100)  compoentNumber <- compoents$ncomp  dist.r <- dist(mainBasins,method = 'euclidean')  #运用欧几里德距离求样本间的距离  fit\_hc <<- hclust(dist.r)  #进行聚类分析  result <<- cutree(fit\_hc,k=k)  最后返回聚类结果。 |

**5.2.3 上传DEM参数文件功能实现**

上传DEM参数文件功能实现过程中涉及到三个类，分别是用来控制DEM参数文件上传过程的UploadTxtServlet，用其中updateDemFlag()来更新Dem状态的EditUtil类和用来解析DEM文件的ExtractDem类。它们之间的关系如下图所示：

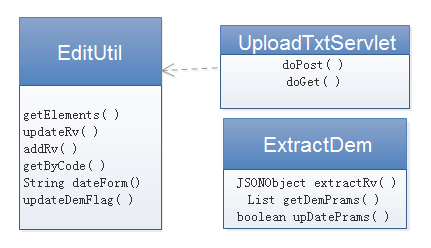


图5.3 DEM文件上传uml类图

首先UploadTxtServlet能从前台表单获取到DEM参数文件在本地的存储路径和它相对应的河流代码这两个参数，当接收完文件后，系统会将它保存在服务器上的一个指定文件夹中，确保文件的安全性，并将文件的名称改成河流代码。实现代码如下：

|  |
| --- |
| public void doGet(HttpServletRequest request, HttpServletResponse response) throws ServletException, IOException {  //上传DEM的河流编码  String rvcode=null;  //得到上传文件的保存目录，将上传的文件存放于WEB-INF目录下，不允许外界直接访问，保证上传文件的安全  String savePath =this.getServletContext().getRealPath("/uploadfiles");  String saveFileName = "temp";  File file = new File(savePath);  //判断上传文件的保存目录是否存在  if (!file.exists() && !file.isDirectory()) {  System.out.println(savePath+"目录不存在，需要创建");  file.mkdir(); //创建目录  } |

系统使用Apache的commons包中的文件上传组件来传输文件：创建一个DiskFileItemFactory工厂；创建一个文件上传解析器；解决上传文件名的中文乱码；使用ServletFileUpload解析器解析上传数据，解析结果返回的是一个List<FileItem>集合，每一个FileItem对应一个Form表单的输入项。代码如下：

|  |
| --- |
| DiskFileItemFactory factory = new DiskFileItemFactory();  ServletFileUpload upload = new ServletFileUpload(factory);  upload.setHeaderEncoding("UTF-8");  List<FileItem> list = upload.parseRequest(request); |

因为上传的DEM参数文件为txt格式，里面除了包含了本文需要的18个DEM流域特征参数，还有其他信息。所以我们要将需要的参数从文本中提取出来放到数据库中，使用的是ExtractDem类。其中上传文件的格式如图所示：

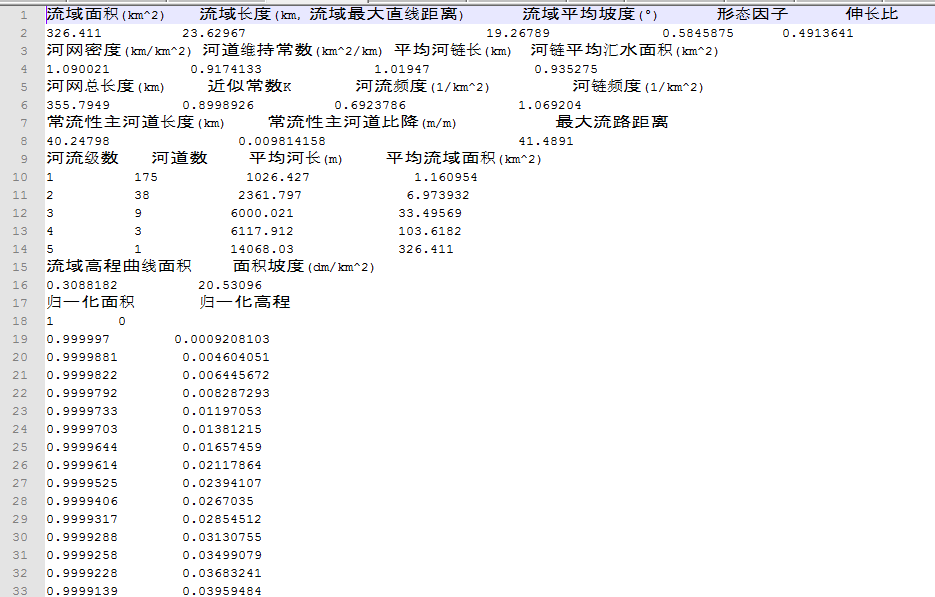


图5.4 文件格式

可以看出，txt格式并不规则，且包含其他特征信息。在ExtractDem类中，按行读取txt文件，直接取得前八行中第二、四、六、八行的值至对应的属性，当行数大于九时，设定一个flag=0，当读到值为“流域高程曲线面积”时，令flag等于该行的数值加一，取flag位置的数值到流域高程曲线面积和面积坡度属性值中。

|  |
| --- |
| JSONObject extractRv(String rvcode,HttpServletRequest request){  JSONObject jsonObject=new JSONObject();  List DemPrams=getDemPrams();  List DemPramsList=new ArrayList();  String message="success";  // 标记“流域高程曲线面积”“面积坡度” 两个参数的位置  if(file.exists()&&file.isFile()){  List lines=file.readLines();  int count=0;  for(String line:lines){  count++;  switch (count){  case 2:  String[] line2=line.split(" ");  for(int i=0;i<line2.length;i++){  if(!line2[i].equals("")){  DemPramsList.add(DemPrams.get(pramIndex)+"="+line2[i]);  pramIndex++;  }  } //获取上传文件目录  String savePath = request.getServletContext().getRealPath("/uploadfiles");  String fileName=rvcode+".txt";  File file = new File(savePath+"/"+fileName);  int pramIndex=0; //下dian面参数顺序  int Flag=0; //  break;  case 4:  String[] line4=line.split(" ");  for(int i=0;i<line4.length;i++){  if(!line4[i].equals("")){  DemPramsList.add(DemPrams.get(pramIndex)+"="+line4[i]);  pramIndex++;  }  }  break;  case 6:  String[] line6=line.split(" ");  for(int i=0;i<line6.length;i++){  if(!line6[i].equals("")){  DemPramsList.add(DemPrams.get(pramIndex)+"="+line6[i]);  pramIndex++;  }  }  break;  case 8:  String[] line8=line.split(" ");  for(int i=0;i<line8.length;i++){  if(!line8[i].equals("")){  DemPramsList.add(DemPrams.get(pramIndex)+"="+line8[i]);  pramIndex++;  }  }  break;  } |

**5.2.4 河流信息管理功能实现**

本系统提供河流信息的新增、更改和查询功能。这三个功能都是依赖DataQueryServlet和EditUtil两个类实现。两个类之间的关系如下所示。

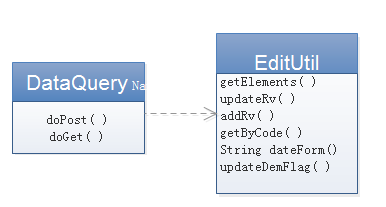


图5.5 UML类图

其中，updateRv()是实现更新河流的方法，addRv()是实现新增河流的方法，getByCode()是实现根据河流编码查询河流信息的方法。

**5.2.5 系统界面实现**

根据上一章的设计，系统界面中的行政区划列表是一个树形结构的选择枝，对全国的城市按省→市→县（区）的级别进行分级显示。通过应用ExtJS框架来实现。ExtJS是一个和后台不相关的，主要作用是新建和美化前端用户界面的框架。它使用javascript语言编写。代码实现如下：

|  |
| --- |
| Ext.onReady(function(){  var addvcdTree = new Ext.tree.TreePanel({  id:'mytree',  renderTo: 'addvcdtree',  title:"行政区划",  height:800,  collapsible:true,  useArrows: true,  autoScroll: true,  containerScroll: false,  border: false,  rootVisible:false,  root: new Ext.tree.AsyncTreeNode({  id:"9999",  text:"根节点",  lines:true,  loader: new Ext.tree.TreeLoader({  dataUrl:"../TreeServlet?method=addvcd&id="+"0"  })  }),  listeners:{  click:function(n){  var addvcd=n.attributes.id;  var text=n.attributes.text;  if(text=='全国'){  addvcd='100000';  }  defaultaddvcd=addvcd;  defaultname=text;  if(analyseflag!=null){  document.getElementById("myiframe").src="analyse.jsp"; }  else{  document.getElementById("myiframe").src=  "prepare.jsp?addvcd="+addvcd+"&title="+encodeURI(encodeURI(text));  }  }  }  });  }); |

**5.3 系统展示**

**5.3.1 系统界面展示**

登录系统后，系统界面如图5.6所示。首页，相似性分析的标识下面是两个选项，单流域搜索和多流域分析，分别点选就能进入各自的相似流域查询功能。在系统右下部分显示数据库中所有的河流信息，左半部分是行政区划表，点选里面的省份或城市，河流信息表就会显示该地的河流信息。河流信息表的上面是新增、更新和上传DEM三个按钮，分别对应新增河流信息、更新河流信息和上传DEM参数文件三个功能。同时还有一个搜索框。

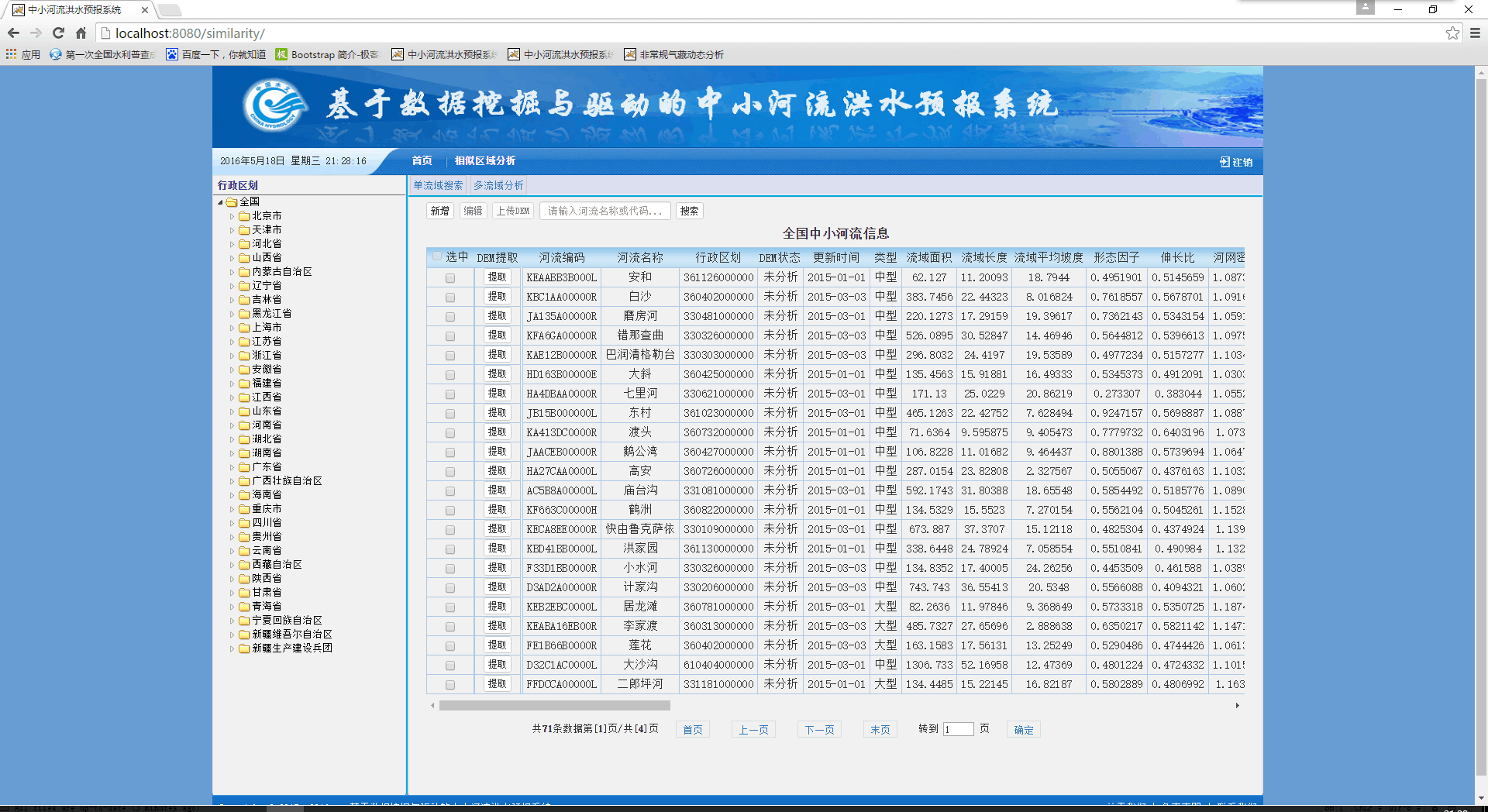


图5.6 系统界面图

**5.3.2 单流域搜索功能**

点击主界面上的单流域搜索功能。在河流信息表中勾选一条作为基准河流的待搜流域，这里以安和为例，勾选安和后，待搜流域下拉框中会出现安和的名字。也可将下拉框下拉，其显示当前页面所有河流的名字，点选其他名字可更换河流。然后勾选河流信息表中其他河流作为搜索范围，可以选不同页的河流，它们的信息会显示到搜索范围的多选框中，在多选框中可以再次确认需要比较的河流。点击重置按钮会清空多选框中的河流；勾选多选框中一部分河流再点击移除时只移除选择的河流。在相似性要素的选栏中勾选需要的分析要素，一般默认全部。最后点击确认，进行分析。如图5.7所示：

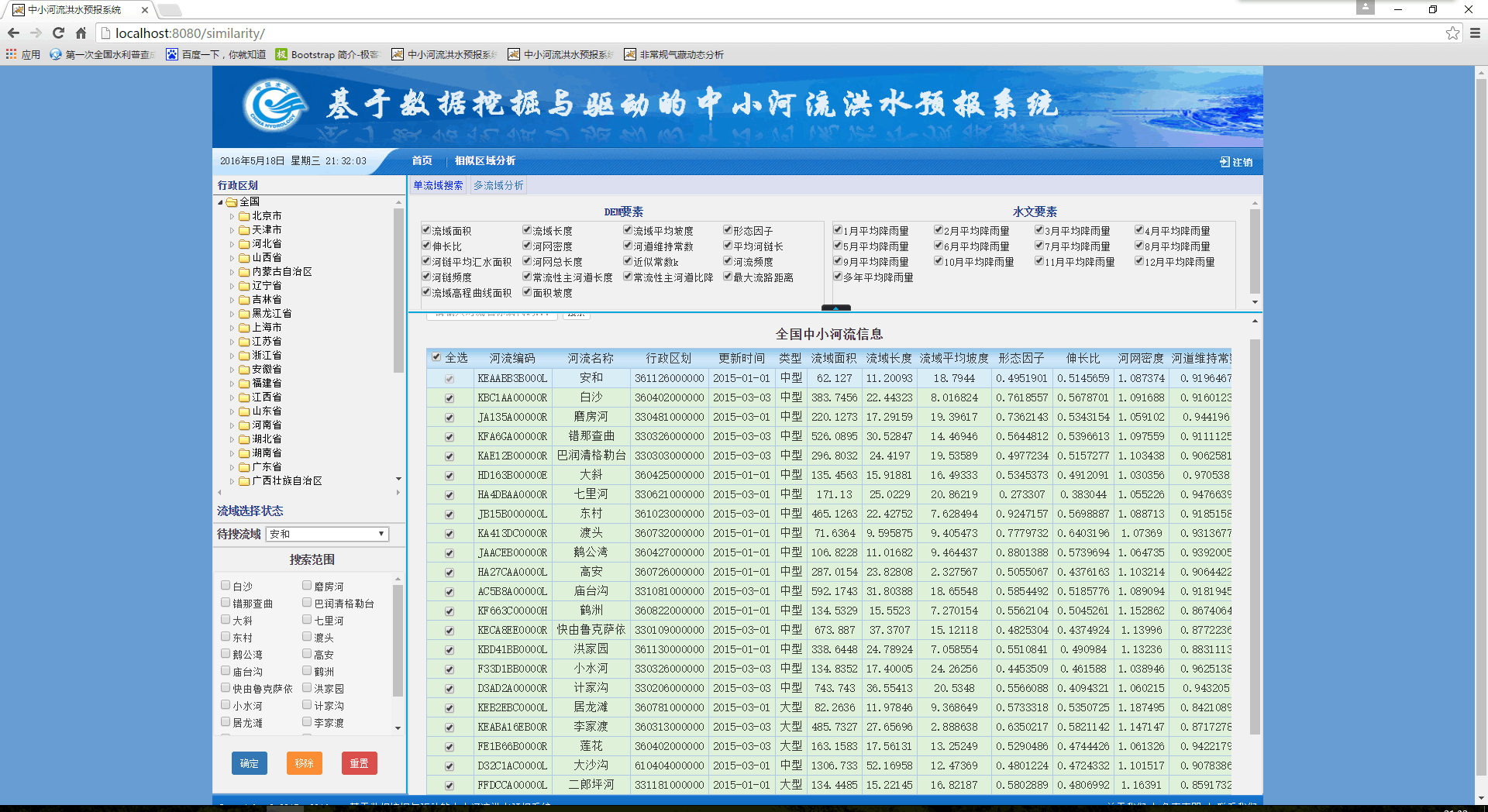


图5.7 单流域搜索

分析完成后结果返回，以聚类图的形式和表格数据的形式进行展示。表格中列出了所有和安和相似的河流，用青色标出。在右上角聚类数目一栏，可以调整用户想要的聚类数目，返回不同结果。如图5.8所示：

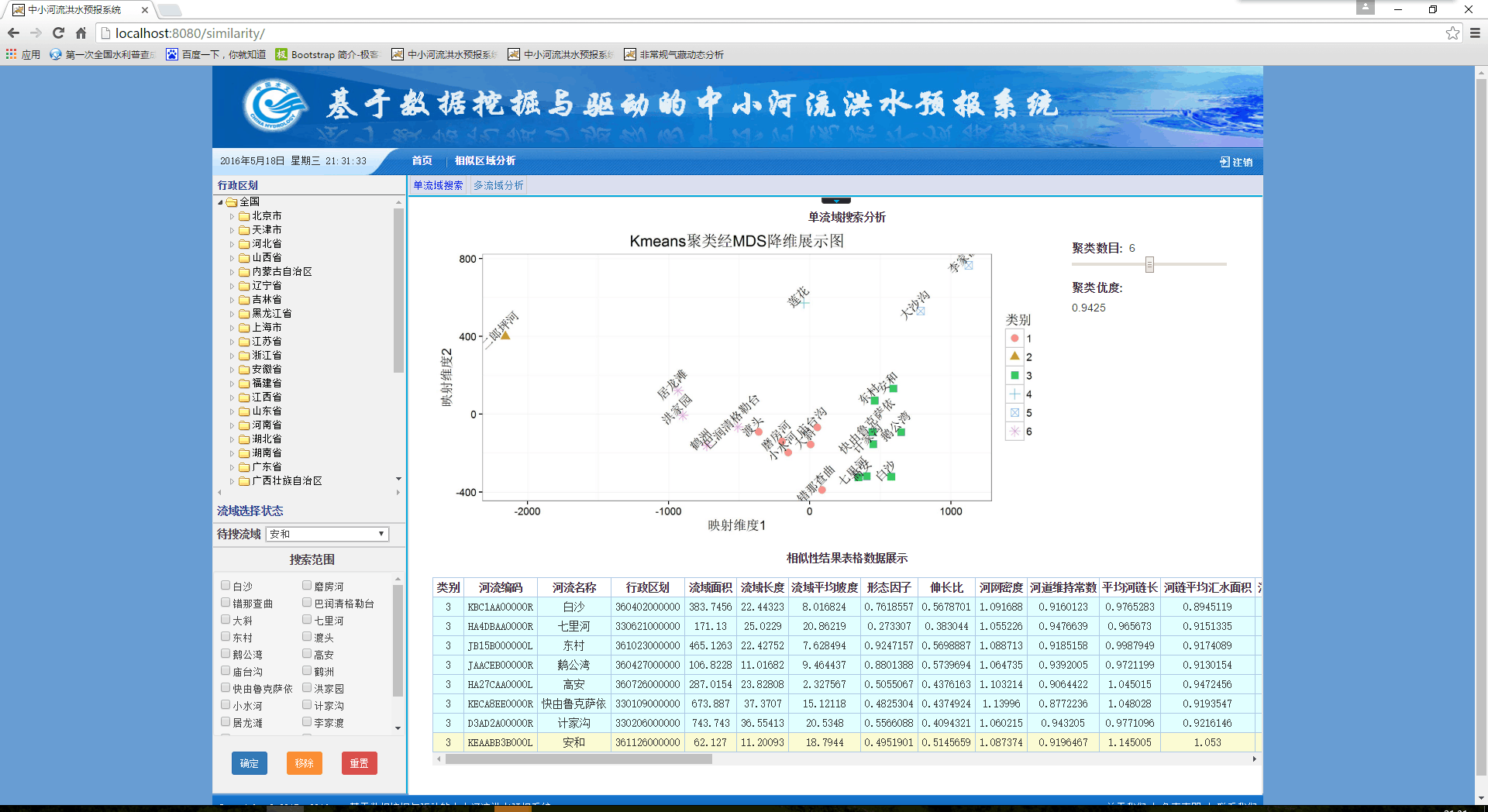


图5.8 单流域搜索结果展示

**5.3.3 多流域分析功能**

与单流域搜索一样在河流信息表中勾选需要的河流，它们会在分析范围中显示。然后在分析范围的多选框中对这些河流进行进一步的筛选操作：勾选不需要的河流，然后点击移除；若全部不需要就点击重置按钮。确认了分析范围后，在相似性要素栏中选择需要的分析要素，一般默认全部要素。最后点击确认进行分析，界面如图5.9所示：

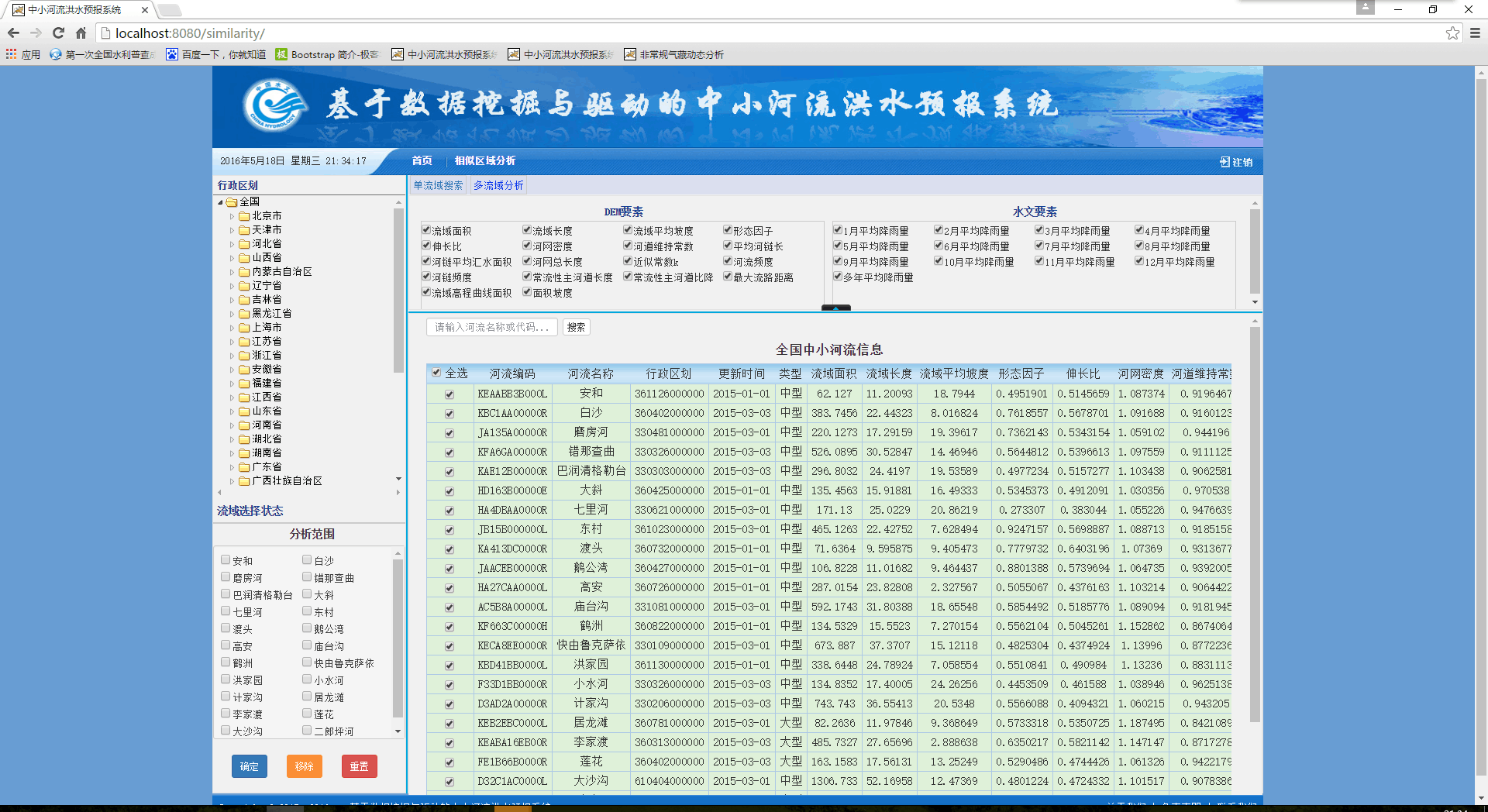


图5.9 多流域分析

分析完成后结果返回，以聚类图的形式和表格数据的形式进行展示，如图5.10所示：

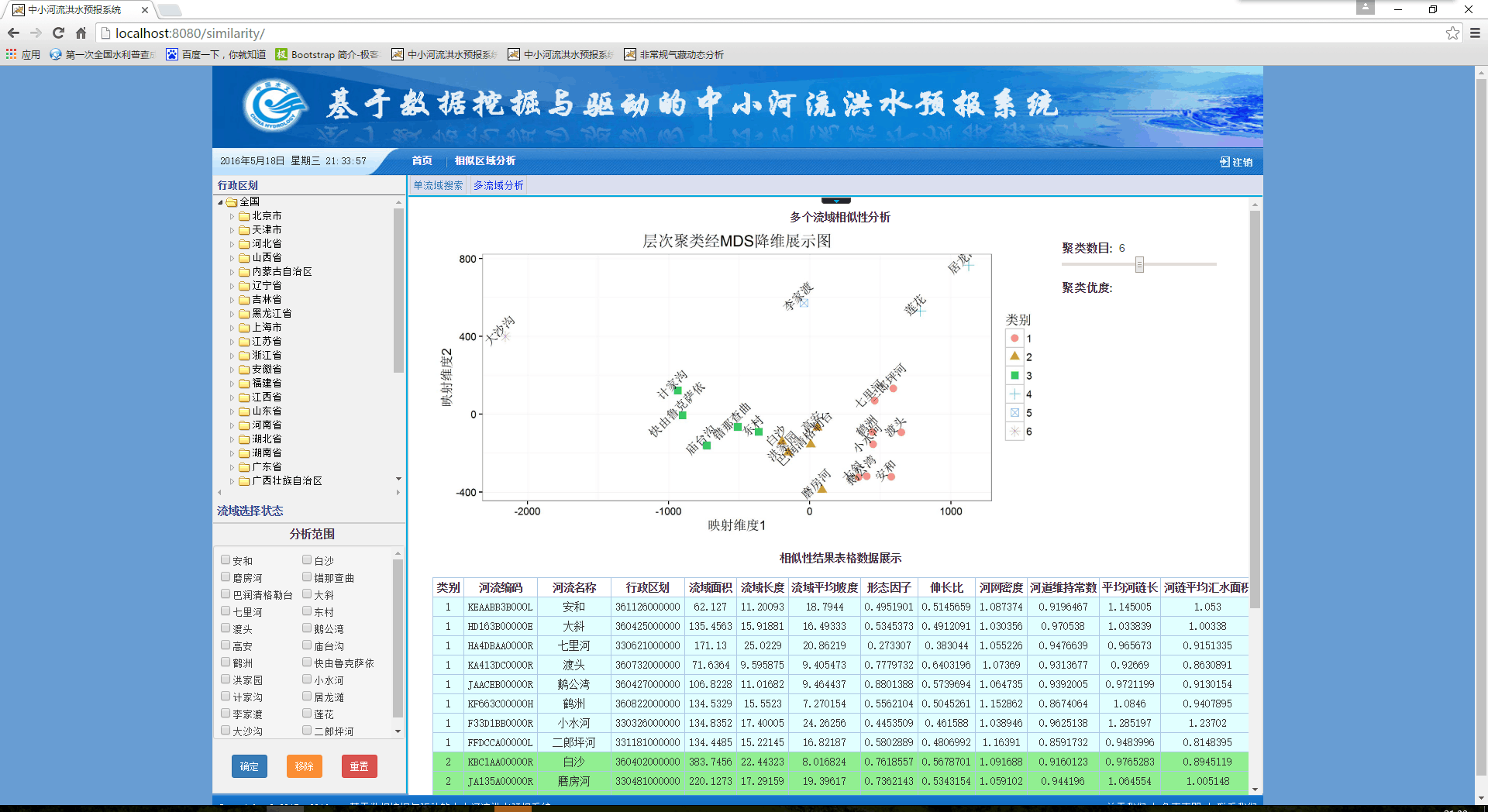


图5.10 多流域分析结果展示

**5.3.4 河流信息查询功能**

在主界面的搜索框中输入河流编码，点击搜索按钮，河流信息表就会显示该河流的信息，如图5.11所示。



图5.11 河流信息查询

**5.3.5 上传DEM参数文件功能**

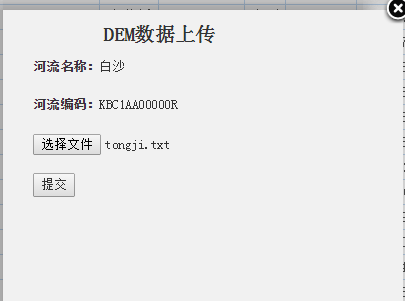
勾选河流信息表中的一条河流，点击上传DEM按钮，会跳出如下界面：

图5.12 上传DEM参数文件

选择上传文件的地址，点击提交，提交成功后会显示“提交成功！DEM状态更新完成。”

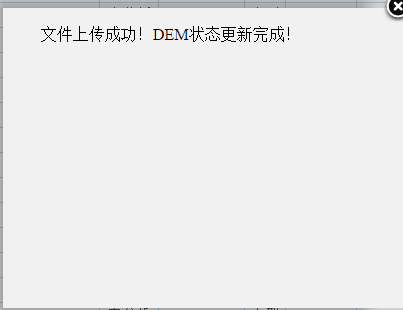


图5.13 上传成功

**5.3.7 新增河流信息功能**

点击新增按钮，在弹出的页面框中填写河流代码、河流名称、行政区划、DEM参数信息和雨量信息，点击提交，其中河流代码、河流名称、行政区划三项不能为空。其界面如下图所示：



图5.14 界面一



图5.15 界面二

**5.3.8 编辑河流信息功能**

勾选一条河流，点击编辑按钮，弹出编辑信息的界面，输入需要修改的值提交即可，页面与新增河流相似。

**5.4 小结**

本章主要介绍了基于DEM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。

**第6章 总结与展望**

流域相似性研究的展开将加强人们对相似流域的理解，根据流域特点进一步对流域类型进行总结，提出研究模型，完善水文资料，提高流域研究的科学性。随着高分辨率遥感卫星、机载传感系统、全球卫星导航系统、合成孔径雷达等对地观测技术的发展, DEM空间数据步入大数据世代。但原始DEM数据并不能直接表达有效的水文信息，运用数据挖掘技术基于DEM数据进行流域特征挖掘对流域相似性研究有重要意义。

本文将数字高程模型引入流域相似性分析中，设计并实现了基于DEM的流域相似性查找系统。对基于DEM流域特征的提取过程进行了详细描述，以及利用相似性分析中的主成份分析、K-Means算法和层次聚类算法在河流特征信息上进行了实现，并运用web技术对查询结果进行展示。

但是，本系统的DEM特征信息还是基于ArcGIS软件手动提取，如何实现相似性指标自动提取和保证河流相似结果的准确性是以后改进的方向。

**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——万定生教授悉心的指导。在每周的毕设讨论会议上万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，万老师都会对我进行督促。万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的沈强、陆宇庆两位师兄，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

[1]蔺彬彬. 基于DEM的三川河流域特征提取研究[D]. 太原理工大学,2013.

[2]刘金涛, 宋慧卿, 王爱花. 水文相似概念与理论发展探析[J]. 水科学进展, 2014, 25(2):288-296.

[3]Tribe A. Automated recognition of valley lines and drainage networks from grid digital elevation models: a review and a new method[J]. Journal of Hydrology, 1995, 167(s 1–4):393-396.

[4]Garbrecht J, Martz L W. Digital Elevation Model Issues In Water Resources Modeling[J]. Geocarto International, 2008, 21(1):21-26.

[5]O'Callaghan J F, Mark D M. The Extraction of Drainage Networks From Digital Elevation Data[J]. Computer Vision Graphics & Image Processing, 1984, 27(3):323-344.]

[7]S. K. Jenson and J. O. Domingue, “Extracting Topographic Structure from Digital Elevation Data for Geographic Information System Analysis,” Photogrammetric Engineering and Remote Sensing, Vol. 54, No. 11, 1988, pp. 1593-1600.

[8]Quinn P , Beven K, Chevalier P , et al . The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models [J ]. Hydrological Processes , 1991 , vol.5 : 59-79.

[9]赵美玲. 基于DEM数据提取水文流域参数分析[J]. 山西水利, 2014(6):46-47.

[10]张振, 宋亚娅. GIS环境下基于 DEM流域水文地理信息的提取[J]. 地下水, 2014(6).

[11]张永红. 基于DEM的辽河流域子流域边界的自动提取[J]. 山东工业技术, 2015(14):262-263.

[12]张国义, 房明惠, 徐云,等. RSI River Tools系统及其应用介绍[J]. 计算机应用, 2002, 22(8):38-40.

[13]Langhaar H L. Dimensional Analysis and Theory of Models[J]. John Wiley & Sons, Inc. New York, N. Y.; Chapman & Hall, Ltd. London, 1951.

[14]RODRIGUEZ-ITRUBE I, VALDES J B. The geomorphologic structure of hydrologic response [J]. Water Resources Research,1979, 15: 1409-1420.

[15]Young A R. Stream flow simulation within UK ungauged catchments using a daily rainfall-runoff model[J]. Journal of Hydrology, 2006, 320(s 1–2):155-172.

[16]丁晶.模糊相似选择在水文计算中的应用[Ｊ].成都科技大学学报,1998,(6).

[17]陈守煜.相似流域选择的模糊集模型与方法[Ｊ].水科学进展,1993,(4).

[18]邓红霞, 李存军, 张少文,等. 基于集对分析的相似流域选择方法[J]. 人民黄河, 2006, 28(7):3-4.

[19]胡余忠, 胡以宝, 章彩霞,等. 基于DEM的流域相似性数值识别与应用[J]. 江淮水利科技, 2010(4).

[20]肖益民, 梅汇海. 相似流域选择的灰关联方法[J]. 武汉水利电力大学学报, 1998(3):10-11.

[21]李亚伟, 陈守煜, 聂相田. 基于PCA和聚类分析的相似流域选择方法[J]. 东北水利水电, 2004, 22(7):1-3.

[22] 赵焱. 基于ArcGIS 10.0的河道地形高程数据处理方法[J]. 电子制作, 2015(5).

[23] <http://wenku.baidu.com/view/8e988c003169a4517723a3f6.html>