泰迪智能科技
TipDM Intelligent Technology

R 语言数据挖掘建模平台

TipDM-RB

技 术 白 皮 书

地址： 广州市黄埔区科学城开泰大道 36 号 1 栋 212
网址： http://www.tipdm.com
邮箱： services@tipdm.com
邮编： 510000
电话： 020-22205718

目录

1. 概述

1.1 背景

近年来，数据挖掘技术得到了飞跃式发展，其能够从大量数据中发现有用的知识，利用这一技术，通过客观统计和分析，可以从大量数据中发现潜在规律，找出隐含的模式，准确掌握未来的动态，因此已经成为各应用领域的重要技术，高校数据挖掘课程的开设也应运而生。但是，数据挖掘课程融合了数据库、信息检索、统计学、计算机编程等多个领域的内容，既包括各种理论知识，又离不开相关的实践技术，如何将理论与实践相结合，培养和提高学生的创新能力及综合解决问题的能力，成为新的难题。为了改变这一情况，R 语言数据挖掘建模平台应运而生。

1.2 产品简介

R 语言数据挖掘建模平台是由广州泰迪智能科技有限公司自主研发，面向高校数据挖掘课程教学的数据挖掘建模工具。平台以实际挖掘案例为切入点，对老师而言，老师在使用平台进行教学时不仅可以讲授数据挖掘基本流程、主要挖掘算法的基本原理，还可以使用平台的示例模板讲解一个算法的应用场景或者是一个完整案例的挖掘步骤，并查看各步骤源代码；老师还可以上传自定义算法，与平台算法进行对比。对学生而言，平台大大降低了学习数据挖掘的门槛，让学生对数据挖掘有了更感性的认识，激发学生的学习兴趣。

1.3 专业术语

系统组件：将建模过程涉及到的输入/输出、数据探索及预处理、建模、模型评估等算法分别进行封装，每一个封装好的算法都可称之为组件。

个人组件：用户可按照平台规定的格式编写脚本，配置相关输入、输出、算法参数，可作为平台组件，反复调用。

工程：为实现某一数据挖掘目标，将各组件通过流程化的方式进行连接，整个数据流程称为一个工程。

模型：主要针对分类、回归算法而言，使用一部分数据用于训练，会得到一个模型，里面将保存算法的参数，可使用该模型对另一批数据进行验证或预测。

任务：支持定时同步数据库数据源至平台或定时运行某一工程。

2. 技术特点

2.1 使用 R 语言计算引擎

平台使用 R 语言作为计算引擎，R 语言主要有以下几个优点：

1. 程序开源，发展成熟，社区在不断壮大，拥有许多开发者；

2. 基于内存，训练速度快，代码量少；

3. 画图精美，具有强大的画图功能；

4. 主要是统计学家为解决数据分析领域的问题而开发的语言，具有强大、完整的统计学及数据分析知识系统。

2.2 使用 B/S 架构

平台使用 JAVA 语言开发，采用 B/S 结构，B/S 结构主要有以下几个优点：

1. 分布性强，客户端零维护，只要有网络、浏览器，就可以随时随地进行查询、浏览等业务处理；

2. 业务扩展简单方便，通过增加网页即可增加服务器功能；

3. 维护简单方便，只需要改变网页，即可实现所有用户的同步更新；

4. 开发简单，共享性强。

2.3 使用工作流形式

平台使用工作流的形式展现数据挖掘的流程，用户可在没有 R 语言编程基础的情况下，通过拖拽的方式进行操作，将数据输入输出、数据预处理、挖掘建模、模型评估等环节通过流程化的方式进行连接，以达到数据分析挖掘的目的。

2.4 提供接口拓展模块

平台提供接口拓展模块，包含所有算法 API（JAR 包）和 WebService 接口，方便用户进行算法优化研究。平台开放灵活的系统对接能力，能帮助用户承接并完成社会项目开发。接口模块基于标准 RESTful 设计，用户可以方便，快捷的通过浏览器在线浏览、测试各个接口。

3. 运行环境

为保证本平台在教学中可正常使用，以 40 个用户并发使用为标准，提供标准运行环境如表 -1 所示。

表 -1 软硬件运行环境说明

	名称	配置要求 （CPU、内存、硬盘）	数量	用途
硬件	服务器	CPU：不小于 2 颗 CPU，8 核 16 线程。建议采用 Xeon E5-2620 v4, 2.1GHz，8T/16C，带宽 6.4GB/s/UNIX，tpmc≧16500 内存：不小于 16G 内存。建议至少采用 2*8GB RDIMM DDR4 2400MT/s 硬盘：不小于 1T。建议采用 SAS 硬盘 1T	2	计算服务器
	服务器	CPU：不小于 2 颗 CPU，8 核 16 线程。建议采用 Xeon E5-2620 v4, 2.1GHz，8T/16C，带宽 6.4GB/s/UNIX，tpmc≧16500 内存：不小于 8G 内存。建议至少采用 2*4GB RDIMM DDR4 2400MT/s 硬盘：不小于 5T 硬盘。建议采用 SAS 硬盘 5T	1	数据服务器

	软件名称	版本号	生产商/来源	用途
软件	CentOS	6.7 以上	The CentOS Project | Legal | Privacy	服务器操作系统
	PostgreSQL	9.5 及以上	The PostgreSQL Global Development Group	数据存储
	Redis	3.0 及以上	Redis Labs	缓存
	R	3.3.3 及以上	The R Foundation	计算引擎
	JDK	1.8 及以上	Oracle	JAVA 运行环境
	Tomcat	8.0 及以上	The Apache Software Foundation	网络服务

网络环境	千兆内网
其它	

4. 产品架构

R 语言数据挖掘建模平台系统架构如图 -1 所示。

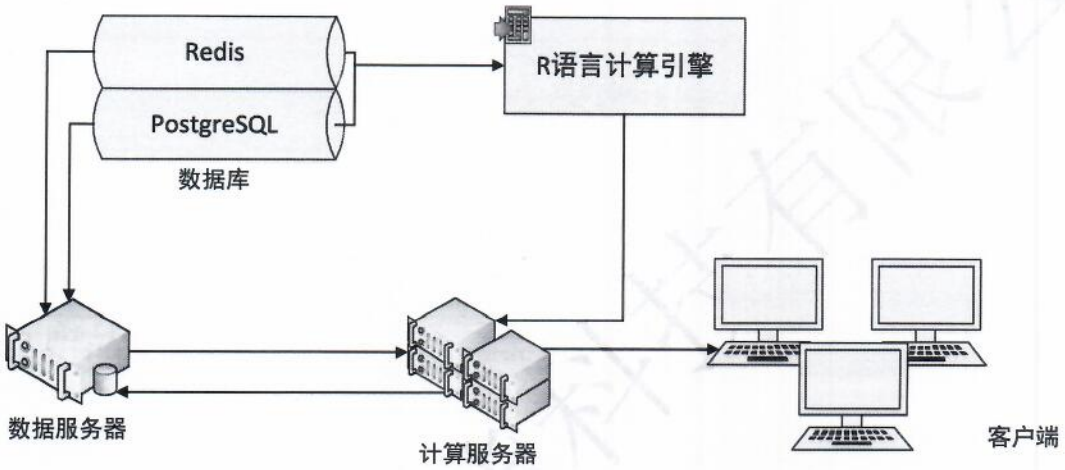

图 -1 系统架构图

平台 Web Server 主要有两大功能：接收前端用户访问请求，对后台数据源或算法调用进行操作。

R 语言数据挖掘平台数据源采用 2 种数据库，分别是 Redis、PostgreSQL。Redis 主要存储系统运行时临时消息，可以高效进行消息的传递及查询。PostgreSQL 主要存储用户数据，当用户数据较少时（这时用户一般运行的是单机算法来对其数据进行处理），使用 PostgreSQL 来存储用户临时数据，同时在工程流程中运行生成的临时表也存储在 PostgreSQL 中。

5. 产品功能

R 语言数据挖掘建模平台主要包括数据源管理模块、组件管理模块、工程管理模块、任务调度、模型管理、系统设置及实现各个具体功能的子模块。

各模块的具体关系如表 -2 所示：

表 -2 产品功能模块关系

模块	功能描述
首页	查看社区、查看入门、通过模板创建工程
数据源	上传数据源、数据分享、数据预览
工程	创建工程、运行工程、参数设置
系统组件	新增系统组件、编辑系统组件、编辑源码
个人组件	新增个人组件、编辑个人组件、编辑源码
模型	导入模型、分享模型、模型预测
任务	新建工程任务、新建数据源任务