**dataminer4web**

队伍：GoldMiner

作者：肖齐天

时间：2013/07/09

# 一.系统概述

## 1.1R语言与weka

r语言擅长于统计分析，weka擅长于数据挖掘。

现在，存储海量数据需要分布式的存储技术和并行式的处理技术，java在这方面有着的太大的优势，而在R语言上做到分布式存储和并行处理，限制很大。

dataminer4web是一款面向web的平台，而不是传统的cs架构。经过多方面的思考，我们决定采用weka作为底层来实现。

## 1.2技术路线

由于采用bs结构和java技术，我们选用开源ajax ui系统dwz作为系统的web前端，采用开源mvc框架eweb4j作为mvc框架。

# 二.系统特性

## 2.1支持多种数据库

本系统支持mysql，mssqlserver，oracle，hsql。

## 2.2多种算法支持

本系统采用weka工具包，支持的算法有

关联分析：apriori

分类：增量式贝叶斯，朴素贝叶斯，J48，SMO

聚类：KMEANS

对于每种算法，可以控制相关参数，反复进行测试

## 2.3海量数据

对于weka而言，要实现海量数据的处理，必须实现增量incremental接口，而增量贝叶斯算法实现了这一接口，从而可以实现海量数据的分类。

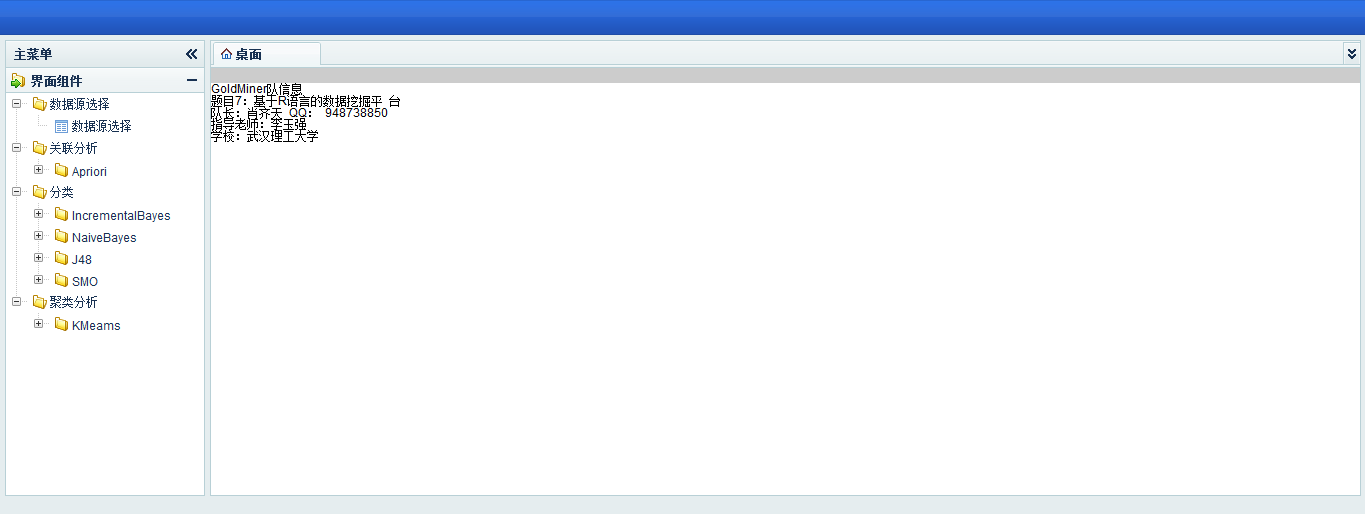
## 2.4缓存处理

为了避免多用户进行重复的数据处理，本系统在web层对数据处理进行缓存管理，而避免每次请求都要重新计算数据。

## 2.5用户界面友好

# 三.系统测试

## 3.1用户界面



### 3.1.1连接数据源

数据库默认为mysql，数据库名为weka，自定义配置请修改weka-stable\_fat.jar中的weka.experiment包中的databaseUtil.props配置文件。

数据采用weather-nominal，天气情况分析

@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}

@attribute temperature {hot, mild, cool}

@attribute humidity {high, normal}

@attribute windy {TRUE, FALSE}

@attribute play {yes, no}

@data

sunny,hot,high,FALSE,no

sunny,hot,high,TRUE,no

overcast,hot,high,FALSE,yes

rainy,mild,high,FALSE,yes

rainy,cool,normal,FALSE,yes

rainy,cool,normal,TRUE,no

overcast,cool,normal,TRUE,yes

sunny,mild,high,FALSE,no

sunny,cool,normal,FALSE,yes

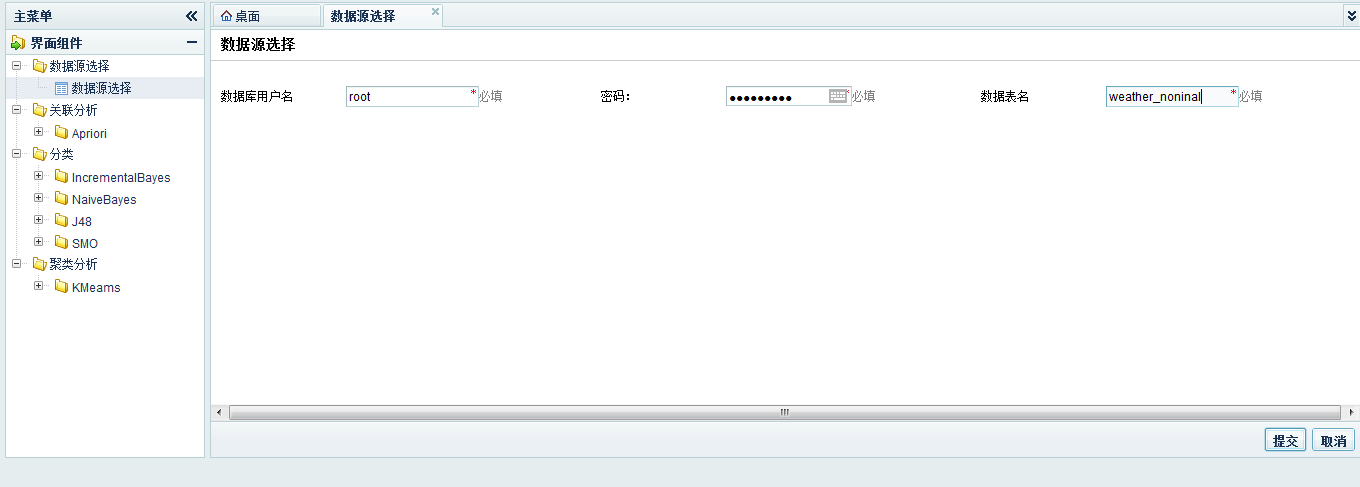
rainy,mild,normal,FALSE,yes

sunny,mild,normal,TRUE,yes

overcast,mild,high,TRUE,yes

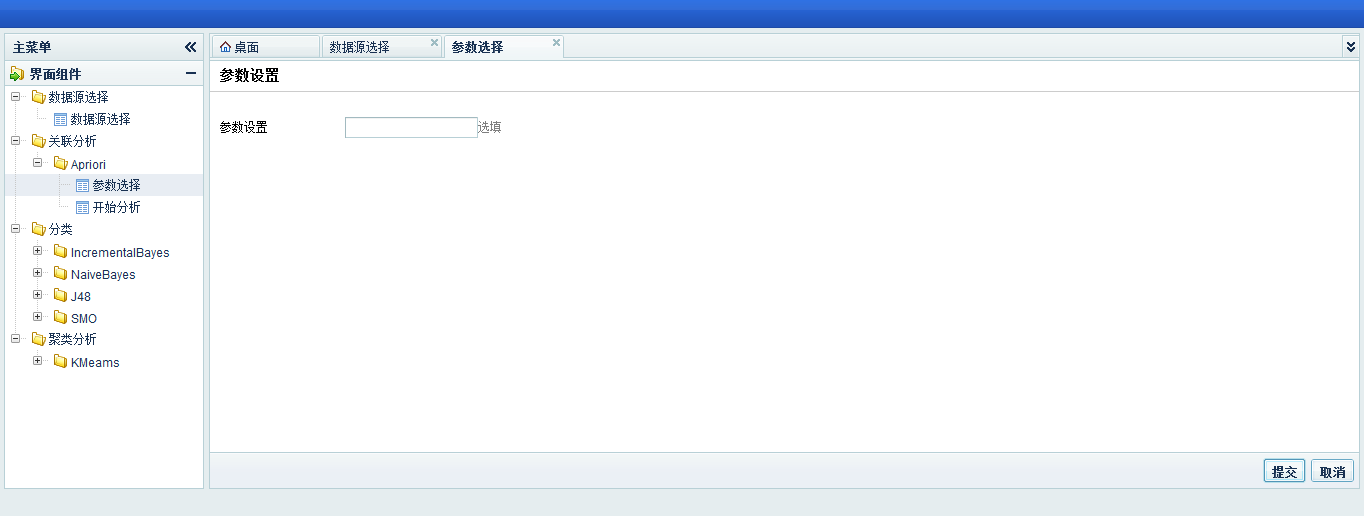
overcast,hot,normal,FALSE,yes

rainy,mild,high,TRUE,no

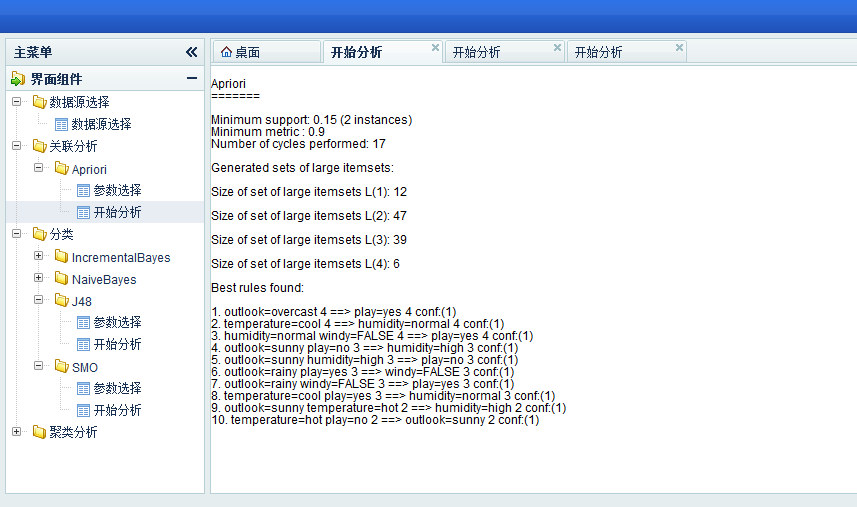


### 3.1.2关联分析

参数设置



关联分析



### 3.1.3分类

J48分类树的建立与验证



### 3.1.4聚类

