Weka代码分析

WEKA作为一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理，分类，回归、聚类、关联规则以及在新的交互式界面上的可视化。

Weka是一个开源的软件，阅读并理解其源代码可以帮助我们更好的理解算法的细节，同时也能够扩展其算法以获得更广泛的应用。但是现在公开的资料中对weka源代码分析的内容很少，所以本博客推出weka源代码分析系列，将针对weka中实现各种算法分析器实现细节。

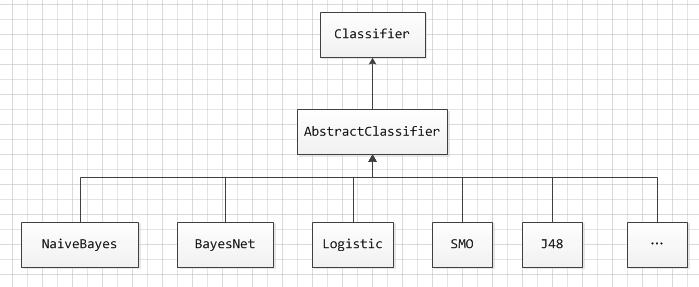
本系列根据weka功能的划分主要分析器分类算法、聚类算法、规则挖掘等，分别介绍其中的各个算法，最后介绍如何对weka进行扩展。本系列主要分析weka3.7.6版本

**1.总述**

本文首先概要列出分类、聚类以及规则挖掘等主要实现的算法以及类结构关系，以便有个大致的了解。

**1.1分类：**

所有的分类器都继承自抽象类AbstractClassifier而AbstractClassifier继承自接口Classifier。集成关系如下图所示：



而类Classifier中主要包含以下接口：

public abstract voidbuildClassifier(Instances data) throws Exception;

该方法主要用于根据样本数据data训练出分类器；

public double classifyInstance(Instance instance) throws Exception;

该方法是对样本instance进行分类并返回最有可能的类；

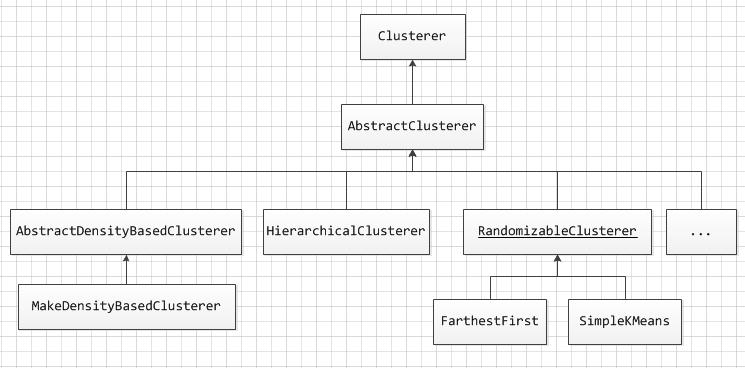
public double[] distributionForInstance(Instance instance) throws Exception;

对样本instance进行分类，如果分类成功则返回的数组值包含该类别，否则数组元素全为0；

所有的classifier必须实现classifyInstance或distributionForInstance。

**1.2聚类：**

所有的聚类算法都集成自AbstractClusterer，而AbstractClusterer继承自接口Clusterer，主要调用关系如下：



Clusterer主要实现的接口如下：

void buildClusterer(Instances data) throws Exception;

根据样本数据训练聚类算法；

int clusterInstance(Instance instance) throws Exception;

public double[]distributionForInstance(Instance instance) throws Exception;

为实例归类；

int numberOfClusters() throws Exception;

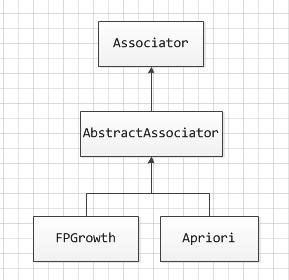
返回聚类个数；

同样所有的聚类算法都必须实现clusterInstance或distributionForInstance。

**1.3规则挖掘：**

关联规则主要实现的算法有Apriori，Fpgrowth等，

类关系如下：



Associator中主要实现

void buildAssociations(Instances data) throws Exception;

该方法主要用于生成规则生成器。