基于WEKA的iris分类Java实现

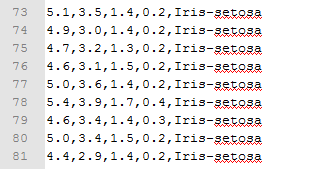
段耀琮，陈刚，韩威

【摘要】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。Iris data set，也称鸢尾花卉数据集。本文讨论了使用WEKA中分类算法对Iris data set进行分类任务的实现。

关 键 词 weka；iris；分类；

1 数据介绍

Iris data set，也称鸢尾花卉数据集，是UCI数据库提供的数据集。如下图Iris data set数据实例，数据集共含150条数据，每条数据有5个属性，属性直接以逗号隔开，中间没有空格。最后一列为本行属性对应的值（分类），即决策属性Iris-setosa。

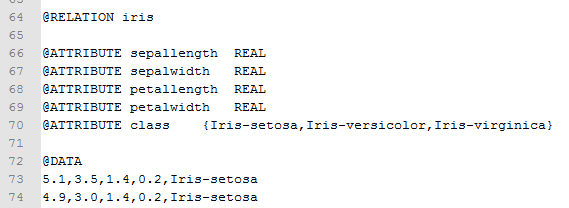


Iris data set数据示例

2 分类实现

2.1 数据处理

WEKA支持的数据格式为.arff文件，该类文件示例如下：



Arff文件示例

其中@RELATION是数据的关系声明；@ATTRIBUTE是数据的属性声明，每一个属性对应一个@ATTRIBUTE语句，来定义它的属性名称和属性类型，并且依次与数据部分的每一列对应；@DATA独占一行，代表数据信息，告诉程序具体数据从下面开始。数据中如有空缺数据则用’？’占位。

对于非Arff文件的数据集需要进行数据转换处理才可以使用。各文件类型间的数据转换可通过数据库系统MySQL及相关应用Navicat for MySQL实现。在我们使用的WEKA3.6版本中，使用DataSource类不止可以读取Arff文件，它同样可以读取其他常见类型的文件，如CSV文件。同时，将数据集导入数据库（MySQL）后，也可以使用InstanceQuery类和JDBC驱动从数据库中读取数据集。我们在此项分类任务中选择了Arff文件数据集和ArffLoader类完成数据的读取工作。

2.2 封装处理

在使用具体算法分类之前需要先进行几步准备工作，如数据读取，属性选择，目标类指定，交叉验证等。

数据读取我们使用WEKA中ArffLoader类，并对常用数据读取操作在MyUtil类中进行简单封装。逻辑代码如下：

ArffLoader loader = new ArffLoader();

loader.setFile(file);

ins = loader.getDataSet();

loader = null;

在数据读取后通常需要对属性进行选择,删除不需要的属性。可以使用WEKA中的Filter类和Remove类完成属性选择工作。对于我们的数据集，不需要这步工作，所以略过。这里将选择属性的逻辑代码给出如下：

Remove rm = new Remove();

rm.setAttributeIndicesArray(attributes);

rm.setInputFormat(instances);

instances = Filter.useFilter(instances,rm);

使用Instances类的setClassIndex方法对目标属性进行指定，这里我们将数据中最后一个属性设置为目标属性：

insTrain.setClassIndex(insTrain.numAttributes() - 1);

其他如交叉验证等工作不是分类任务必须的，这里不多论述。

2.3 分类和结果

选择不同的算法和训练集对分类器进行训练，需要用到WEKA中Classifiers类。同时可以使用Evaluating类和交叉验证对分类器评估。最后讲分类器结果和评估结果输出，根据输出结果可以挑选合适的算法训练分类器。为了详细观察分类器分类情况，我们使用outputResult方法将分类结果进行对比打印。

部分代码和结果给出如下：

Public static void outputResult(Instances insTrain,Instances insTest,Classifier cfs) throws Exception{

int num = insTest.numInstances();

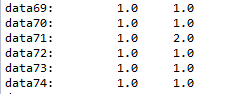
for(int i=0;i<num;i++){

double predicted = cfs.classifyInstance(insTest.instance(i));

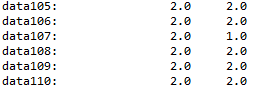
System.out.println("data"+(i+1)+":\t"+"\t"+insTrain.instance(i).classValue()+"\t"+predicted);

}

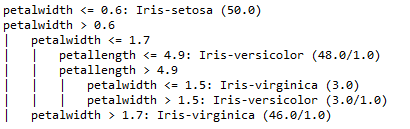
}



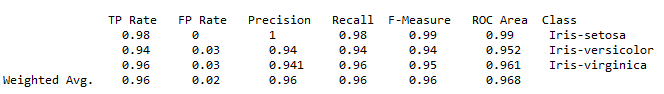
分类结果对比-错误分类1



分类结果对比-错误分类2



J48算法分类树



分类详细信息

3 总结

J48分类算法训练出的分类器具有良好的分类性能，可以较好地完成Iris数据集的分类任务。除分类外，WEKA也提供了大量优秀的聚类、关联规则等其他数据挖掘算法，恰当地使用可以方便快捷的完成多种数据挖掘任务。

参考文献

[1]