数据挖掘大作业实验报告

互联网广告分类器设计与训练

胡玮玮(2006011293) 孔祥欣(2006011299) 徐恺(2006011281)

清华大学计算机科学与技术系计62班

{huww06, kxx006, jaguarxk000}@gmail.com

摘要：

## 0 引言

近年来随着互联网的不断发展，产生了各种各样的电子数据，如何在海量的数据中挖掘出有用的信息成为了一个新的课题。在这个背景下，我们学习了数据挖掘课程，并利用所学的知识设计与实现一个基于文本的分类器，旨在能够对符合一定特征的数据进行分类。

## 1 实验内容

数据挖掘课程提出了互联网广告分类器设计与训练项目。项目提供了一个数据集，其中包括了在互联网页面上呈现出的可能广告，并提供了若干属性，对于离散的属性值来说只有0和1两种分类结果。还包括了三个连续的属性，包括图片长度、宽度、比例等。除此以外提供了一些数据，针对每一个记录（instance）将其定义为广告（ad）或者非广告（nonad）。

项目要求从这些提供的数据中挖掘出适当的有用信息，以这些有用的信息作为依据，对其他给定的若干数据进行分类，判定是否属于广告。

项目建议使用weka工具进行实验，同时鼓励自己编程实现数据挖掘的过程，或者对相关算法进行适当的改进。

## 2 算法分析

我们首先对这个项目的要求进行分析，了解项目的目的，以及提供的数据集的格式。进而熟悉数据挖掘的流程，经过课堂的学习和查阅相关论文，我们认识如下：

数据挖掘过程中存在如下两个名词，一个是训练集，一个是测试集以及相应的标准答案，我们需要做的事情是从训练集中训练出一个分类器，进而使用这个分类器对测试集中的数据进行测试，与标准答案比较即可得到该分类器的效果。一个典型的分类过程包括如下几个步骤：数据预处理→特征选择→训练分类器→分类。因此我们将从以上的四个部分来对我们的算法进行介绍。

### 2.1 数据预处理

在进行数据预处理之前，我们首先需要认识项目已经提供给我们的数据文件格式。我们并没有使用助教已经构造好的ad.arff文件，相反我们使用了ad.names和ad.data两个文件。下面进行分别介绍：

ad.name文件：该文件的主要作用是提供了数据的若干属性，其中一条典型的语句如下格式：

url\*likesbooks.com: 0,1.

其中“\*”号之前的“url”表示该属性的类别是一个url，“\*”号与“：”之间的likesbooks.com表示该属性的名称，而“：”以后的值以“，”为间隔分别表示该属性的取值范围。

ad.data文件：该文件的主要作用是提供了所有的相关数据，其中一条典型的语句如下格式：

125,125,1.0,1,0+,1+ ,ad.

假设在ad.name文件中属性个数为n，那么在以上的一条数据中，以“，”为分隔符，应该可以得到n+1个值，其中前n个值分别对应每一个属性的取值，最后一个值标记该条记录是否是广告，“ad”表示该记录为广告，“nonad”表示该记录非广告。

了解完了文件格式以后，下面我们来进行数据处理，在导入文件的过程中，我们进行了如下处理：

首先读取ad.name，我们定义了一个Vector<String> attribute;用来保存所有属性，读完以后得到一个属性数目attributes。进而读取ad.data，我们定义了Vector<String> dataString;来保存原始的数据，读完以后得到一个记录条数instances。

然后我们针对dataString中的内容进行处理，对于其中的每一条记录，首先根据后缀判断该条记录是否是广告，并将该记录保存到Vector<Integer> type;中。然后通过“，”分词，当值为“？”即表示该值缺失，那么将其设置为“0”，由于前三列为连续值属性，所以对其进行离散化，采用在最大与最小值之间进行均分的方式。当这些都处理完毕以后将数据保存至int [][]data中。到此为止，数据处理结束，已经给下一步骤提供了比较符合规范的数据。

### 2.2 特征选择

以上数据经过处理以后虽然比较规范，但是可以发现它们具有很高的维度，如果以这样的训练集来训练分类器的话那么对大大影响训练的效率。因此需要对数据进行降维操作。传统的特征选择算法有特征频率、期望交叉熵、信息增益、χ2统计量等方法。在本次实验中我们实现的算法有特征频率TF和χ2统计量CHI算法。下面进行详细介绍。

2.2.1 特征频率TF

特征频率（Term Frequency, TF）指文本集中特征词t出现的次数，是最简单的特征选择方法。它认为特征在文本集中出现次数越多，对文本分类的贡献越大。在本次实验中，通过int []frequency数组来对各个特征词出现的次数进行统计，设置域值TFYIELD，当特征词出现次数大于该域值时选择此特征项。本实验中选择Vector<Integer>[] processedData;来保存经过特征选择以后的数据，Vector<Integer> result;用来保存经过选择以后的特征词的序号。

2.2.2 χ2统计量CHI

χ2统计量（Chi-square, CHI）的主要思想是：认为词条与类别之间没有独立性，并可类比为一个自由度的χ2分布，χ2统计量的值越高，词条和类别之间的独立性越小，相关性也就越强。特征词t对类别ci的CHI值公式如下：

其中A表示属于ci类且包含t的文档次数，B表示不属于ci类但包含t的文档次数，C表示属于ci类但不包含t的文档次数，D表示不属于ci类也不包含t的文档次数，N为训练语料中的文档总数，且N=A+B+C+D，同时满足要求A\*D>B\*C。分别计算t对每一类的CHI值，再用下式计算词条t对于整个训练集的CHI值。

其中c表示总类别数，P（ci）表示样本属于ci类的频率。在本次实验中设置域值CHIYIELD，根据以上公式计算的结果，过滤低于该域值的词条，保留高于该域值的词条作为特征项，然后对数据的维数进行降维即可。

### 2.3 训练分类器

文本分类系统就是按照文本的主题以及之前选择好的特征值，将具体的文本划归为适当类别的计算机系统，该系统的一个核心部分就是现在所说的分类器。分类器是针对已经降维以后的训练集，采用适当的分类算法，得到一个分类函数。该分类函数输入一个测试数据以后能够返回是否属于某一类。在这个过程中，采用的分类算法是一个关键，常用的分类算法主要分为两类：一是基于统计的方法，如简单贝叶斯、k最近邻方法、类中心向量方法、回归模型、支持向量机等，还有一种是基于规则的方法，如决策树、FOIL等。在本次实验中，我们采用的是FOIL算法以及对FOIL算法改进以后的PRM算法，具体算法设计如下：

2.3.1 FOIL算法

2.3.2 PRM算法

### 2.4 分类

分类是本次挖掘过程的最后一步，针对训练集采用分类算法训练以后得到分类器，实际上分类器保存的就是很多规则，根据这些规则可以判断某数据是否属于某类。对测试数据进行一定的处理，主要是降维工作，然后将处理之后的数据作为参数传入到分类器中，分类器即可返回一个测试结果。将此测试结果与标准答案集相比较，经过适当计算，可以得到本次挖掘过程的准确率与覆盖率。

## 3 实验过程

本次实验中助教推荐了weka这个工具，但是我们并没有使用，我们是自己编程实现了整个挖掘过程，包括数据预处理、特征选择、训练分类器、分类等步骤。

### 3.1 实验环境

### 3.2 实验系统搭建

### 3.3 实验测试

## 4 实验结果

## 5 实验总结

## 参考文献

[1] Xiaoxin Yin and Jiawei Han, CPAR: Classification based on Predictive Association Rules;

[2] 高亚波，文本分类系统的设计与实现，北京交通大学专业硕士学位论文，2008年6月；

[3] 张俊丽，文本分类中的关键技术研究，华中师范大学硕士学位论文，2008年5月；

[4] 余俊英，文本分类中特征选择方法的研究，江西师范大学硕士学位论文，2007年5月；