一种基于 C4.5 决策树的 Web 页面分类算法①

曹 薇[1] 张乃洲[2]

(1.武汉职业技术学院 计算机学院 湖北 武汉 430074；2.湖北大学 知行学院 湖北 武汉 430011)

摘 要： WEB 文本自动分类在很多方面都有着重要的应用，如信息检索，新闻分类等。决策树算法是一种简单并且广泛使用的分类方法，具有很多优点如：分类精度高，分类速度快等。主要研究了运用 C4.5 决策树构建 Web 页面分类器的基本方法和过程,并提出了一个基于 C4.5 决策树的 Web 页面分类器的框架。在此基础上实现了一个运用于网络爬虫的 Web 页面分类器，实验结果表明该算法是非常有效的。

关键词： WEB 文本分类; C4.5 决策树; 信息论; 信息增益率; 网络爬虫

A C4.5 Decision Tree Based Algorithm for Web Pages Categorization

CAO Wei[1], ZHANG Nai-Zhou[2]

(1.Computer College, Wuhan Institute of Technology, Wuhan 430074, China; 2. Zhixing College, Hubei University, Wuhan 430011, China)

Abstract: Web text categorization can be applied to many domains such as information retrieval, news categorization, etc. Decision tree algorithm is a simple method for categorization and has been used extensively. This paper investigates the basic method and process to build a web classifier by means of C4.5 decision tree, which has various merits such as high categorization precision, high categorization speed, etc. Moreover, this paper proposes a C4.5 decision tree based frame of web pages classifier, and implements it on a web crawler. The experimental results show that this algorithm is highly effective.

Keywords: web text categorization; C4.5 decision tree; information theory; information gain ratio; web crawler

1 引言

目前，Internet 呈现出指数式的发展趋势，WEB 为用户提供了海量的数据资源。如何从海量的 WEB 数据源中寻找用户需要的数据，是目前 WEB 信息检索领域的研究热点。而 WEB 文本自动分类问题是 WEB 信息检索领域的一个基本问题。很多相关的研究都可以归结为分类问题，如在信息检索、产品和新闻分类等方面，WEB 文本自动分类都有重要应用[1-3]。

WEB 文本自动分类问题可以归结为如何构建一个高效分类器的问题。目前构建分类器的方法主要有贝叶斯(Bayes)分类算法、k 近邻(KNN)算法、线性最小二乘法估计(LLSF)、支持向量机(SVM)、决策树算法等。在这些分类算法中，决策树算法是一种简单并且广泛使用的分类方法。

决策树算法最早是在 20 世纪 50 年代由亨特在 CLS(Concept Learning System)中提出，后经发展由 J.R Quinlan 在 1979 年提出了著名的 ID3 算法。ID3 算法是以信息论为基础，以信息熵和信息增益(Information Gain)为衡量标准，从而实现对数据的归纳分类，其特点主要针对离散型属性数据[4]。C4.5 算法是 ID3 的改进算法，它在 ID3 基础上增加了对连续属性的离散化处理，并且采用了信息增益率(Gain

①收稿时间:2010-03-07;收到修改稿时间:2010-04-09

Ratio)作为分类的标准，从而克服了 ID3 在用信息增益选择属性时，偏向选择取值多的属性的不足[4-7]。

决策树算法主要具有以下优点[5, 6]：

1) 分类精度高，并以信息论作为理论基础。

2) 决策树的构建所需的耗费较低。即使训练集的规模较大，也能快速构建决策树。

3) 决策树的分类过程速度很快，其分类的时间复杂度为。其中，h 为树的高度。

4) 对于噪声数据具有很好的抗干扰性。

本文介绍了 C4.5 决策树算法的基本原理和应用过程，并给出了一个在实际的 Blog 信息检索系统中，如何构造基于 C4.5 决策树的分类器的例子。在该信息检索系统中，网络爬虫在爬取 Web 页面时，可以利用该分类器来判断页面的类型。实验结果表明该方法是非常有效的。

2 WEB页面分类的基本框架

2.1 Web 页面分类的基本架构

图 1 给出了 Web 页面分类的一般框架。整个框架分为训练过程(学习过程)和分类过程。其中特征提取过程非常关键，Web 页面的特征表示和提取将直接影响到分类的精度。在图 1 中，如果使用不同的模型如决策树、贝叶斯方法或 KNN 等，那么系统可以对应于不同的分类器。本文使用的是 C4.5 决策树模型。

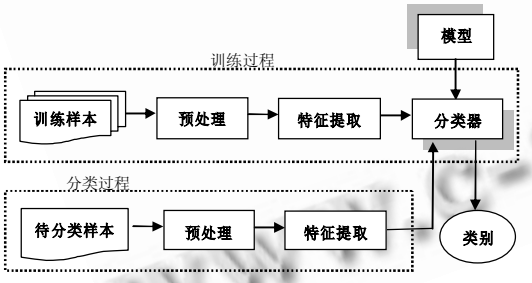

图 1　Web 页面分类的基本架构

2.2 评价指标

评价分类器的指标主要有以下三个[5, 8]：

1) 精度(Precision)

$$Precision = \frac{t\_pos}{t\_pos + f\_pos} \quad (1)$$

其中 *t_pos* 为被正确分类的正例的数量，而 *f_pos* 为被错误分类的正例(即被分类为反例)的数量。

2)召回率(Recall)

$$\mathrm{Re}call = \frac{t\_pos}{t\_pos + f\_neg} \quad (2)$$

其中 *t_pos* 为被正确分类的正例的数量，而 *f_neg* 为被错误分类的反例(即被分类为正例)的数量。

3) 综合指标 F-measure

$$F = \frac{2 * Precision * \mathrm{Re}call}{Precision + \mathrm{Re}call} \quad (3)$$

3 C4.5决策树算法

3.1 C4.5 决策树算法

C4.5 决策树的主要算法思想是[4-6]：

1) 对当前训练集中的各个特征属性 A_i，分别计算出其各自的信息增益率(Gain Ratio)。

2) 选择信息增益率最大的属性 A_k 作为当前决策树的根。

3) 在 A_k 处，将 A_k 属性值相同的例子归为同一子集。

4) 对每一子集，若当前集合中既包含正例又包含反例，则递归调用该算法。

5) 若当前集合中只包含正例或反例，则表明该子集为叶节点，将对应分支标注 P 或 N，然后返回调用处。

3.2 信息增益率（Gain Ratio）的计算方法

下面给出 C4.5 决策树算法中，用于计算信息增益率的主要计算公式[4-6]：

1) 类别的信息熵：

$$H(C) = -\sum_j P(C_j)\log_2(P(C_j)) \quad (4)$$

其中 C_i 为类别 C 的一个取值。

2) 类别条件的信息熵：

按属性 A 将数据集合 D 分割后，其类别条件的信息熵为：

$$H(C \mid A) = -\sum_j \sum_i P(a_i)P(C_j \mid a_i)\log_2(P(C_j \mid a_i)) \quad (5)$$

其中 a_i 为属性 A 的一个取值。

3) 信息增益 Gain(互信息)：

$$I(C, A) = H(C) - H(C \mid A) \quad (6)$$

4) 属性 A 的信息熵：

$$H(A) = -\sum_j P(a_j)\log_2(P(a_j)) \quad (7)$$

5) 信息增益率(Gain Ratio)：

$$gain\_ratio = I(C,A)/H(A) \quad (8)$$

在实际的计算中，相应的概率计算可以用训练集中相应的统计值予以近似。其具体的计算过程如下[4]：

1) 设 D 为训练集，特征属性集合为{A_1，A_2，.. A_m}，类别集合为{C_1，C_2，.. C_k}。现在选择一个属性 A_i，设 A_i 有不相交的属性值集合{ a_1, a_2，.. a_n}。那么 A_i 可将 D 划分为多个子集 D_1, D_2,…. D_n。且 D_i 中对应的属性 A_i 所有的属性值均取 a_i。

2) 设|D|为训练集的样本数；$|D_{a(i)}|$为 D 中，$a=a_i$ 的样本数；$|C_j|$为整个训练集中 C_j 类的样本数；$|C_{ja(i)}|$为 D 中，$a=a_i$ 的样本中，属于 C_j 类的样本数。

根据以上的假设，则有如下的计算公式：

$$P(C_j) = \frac{|C_j|}{|D|} \quad (9)$$

$$P(a_i) = \frac{|D_{a(i)}|}{|D|} \quad (10)$$

$$P(C_j \mid a_i) = \frac{|C_{ja(i)}|}{|D_{a(i)}|} \quad (11)$$

根据以上的三个公式，很容易计算出各属性节点的信息增益率。

4 基于C4.5决策树的页面分类算法及实验

4.1 Web 页面分类问题及特征表示

在很多 Web 应用中，如 Web 信息检索、Web 信息提取等，有时经常需要判断页面的类型。例如在网络爬虫爬行页面时，判断当前页面是否需要爬行的一个特征之一，是判断该页面的类型。在此场景下，我们把页面类型分为 Link Page 和 Detail Page。Link Page 的特点是页面具有较多的链接，适合 Crawler 进行爬取。Detail Page 的特点以文本为主，适合 Web 文本的提取。该问题看起来很简单，似乎只需要统计页面的链接数即可判断，但实际上并非如此。因为目前的 Detail Page 出于商业等目的，往往也包含许多与主题无关的链接，因此需要考虑更复杂的情况。根据观察，我们制定了一些启发式规则，来描述这两种类型的页面的特征：

① Pagetype(两值属性)，取值为 Y 或 N，表示该页面的 URL 是否以 html、htm、shtm 结尾。一般来说，Detail Page 往往取 Y。

② TextDegree(连续数值型)，TextDegree= log2(plainTexts/k)。plainTexts 代表整个文档中去除链接文字后的文本数，k 为一经验常数，其含义是对于一些 plainTexts<k 的页面，表明其文本度较低。TextDegree 越大，该页面属于 Detail Page 的可能性越大。

③ LinkDegree(连续数值型)，LinkDegree = log2(TotalWords / linknums)。TotalWords 代表整个文档中的文本数，linknums 表示该文档中的链接数目。LinkDegree 越小，该页面属于 Link Page 的可能性越大。

④ aggregation(连续数值型)，代表 Web 页面中，文字的聚集度，一般 Detail Page 的显著特征是文本以块(block)的方式出现。因此，aggregation 越大，该页面属于 Detail Page 的可能性越大。该特征主要是通过计算整个文档中文本块(block)的大小获得的。

⑤ catalog(L/D)，类别标记。L: Link Page，D: Detail Page。在训练集上，该标记由手工标注。

根据以上的四个特征，我们在数据集 WebSet10 上对所有的 Web 页面进行了特征表示和提取。WebSet10 是我们从 10 个知名的站点如新浪、搜狐等搜集的 Web 页面集合。该集合一共包含了 1625 个 Web 页面，其中 Link Page 类型的页面为 830 个，Detail Page 类型的页面为 795 个。表 1 是对该数据集进行特征提取后，对应的特征数据集合的片段。

4.2 C4.5 分类器及实验结果

在确定了 Web 页面的特征表示之后，根据第 3 节给出的 C4.5 决策树算法，我们设计了基于 C4.5 决策树算法的 Web 页面分类器。构建的决策树，如图 2 所示，树的深度为 4。我们在前面提到的 WebSet10 数据集上采用了 5 折交叉验证的方法(5-fold cross validation)[5]进行分类训练和测试。其具体的过程如下：将整个数据集分成 5 个大小相等的不相交子集，然后进行 5 轮训练和测试。每次使用 4 个子集作为训

练集，剩下的 1 个子集作为测试集，下一次进行轮换。这样，每个子集都有机会作为一次测试集。最后取其平均的精度、召回率及 F-measure。实验所得的测试结果为：

Precision=0.8722

Recall=0.9336

F-measure=0.9018

表 1　特征数据集片段

TextDegree	aggregation	Pagetype	LinkDegree	catalog
0.1375035	-0.74761283	N	-2.2920439	L
-3.643856	-1.7369655	N	-1.6896598	L
-3.643856	-1.7369655	N	-1.6896598	L
3.7805727	-0.91566855	Y	0.57085675	D
2.1984942	0.64104927	Y	0.87578005	D
4.3575521	-0.55939692	Y	1.7724136	D
3.1501534	-1.5963	Y	-0.3096845	L
2.9140861	-1.8814974	N	1.2697513	D
2.9140861	-1.8814974	Y	1.2697513	D
1.6028844	0.05829864	N	-0.3524998	D
1.7015491	-0.21190004	Y	-0.2537566	D
3.6633446	-0.5076254	Y	0.51331282	D
2.0669503	1.4354306	Y	0.25873426	D
2.0992951	0.84730923	Y	0.29109833	D
1.7865964	1.3465902	N	-7.24E-02	D
0.8678964	-1.7544531	Y	-0.9414361	L

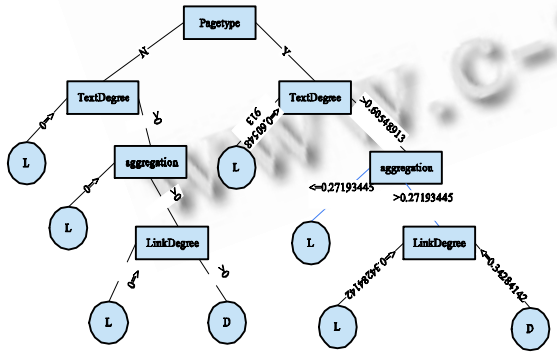

图 2　生成的决策树

分类的结果是令人满意的。在我们设计的 Blog 信息检索系统中，该分类器被运用到了网络爬虫中，实际运行效果非常好。

5　结语

本文主要研究了运用 C4.5 决策树构建 Web 页面分类器的基本方法和过程，并给出了一个基本的算法框架。在此基础上，我们实现了一个运用于网络爬虫的基于 C4.5 决策树的 Web 页面分类器，实验结果表明该方法是非常有效的，该分类器有着较好的应用前景。

参考文献

1 李净,袁小华,沈晓晶. Web 网页信息文本分类的研究. 计算机工程与设计, 2008,29(23):6026－6028.

2 苏金树,张博锋,徐昕.基于机器学习的文本分类技术研究进展.软件学报, 2006,17(9):1849－1854.

3 张翔,周明全,耿国华,侯凡.面向中文文本分类的 C4.5Bagging 算法研究.计算机工程与应用, 2009,45(26):135－137.

4 陈文才,黄金才.数据仓库与数据挖掘.北京:人民邮电出版社, 2004.

5 Han JW, Kamber M.数据挖掘概念与技术.北京:机械工业出版社, 2004.

6 Tan PN, Steinbach M, Kumar V. 数据挖掘导论.北京:人民邮电出版社, 2006.

7 冯少荣,肖文俊.基于样本选取的决策树改进算法.西南交通大学学报, 2009,44(5):643－647.

8 Wikipedia. Precision and recall[2010-02-25]. http://en.wikipedia.org/wiki/Precision_and_recall, 2010-02-23.