第 23 卷 第 6 期
2005 年 12 月

石河子大学学报(自然科学版)
Journal of Shihezi University(Natural Science)

Vol. 23 No. 6
Dec. 2005

文章编号:1007-7383(2005)06-0769-03

三种文本分类算法的比较

王 潇,胡 鑫
(西北师范大学数学与信息科学学院,甘肃兰州 730070)

摘要:介绍了一种基于贝叶斯定理的文本分类模型"树桩网络(stump network)"。将该方法与朴素贝叶斯文本分类器和 TAN(tree augmented naive bayes)文本分类器进行实验比较。结果表明,在大多数数据集上该文本分类方法具有较好的分类正确率。

关键词:计算机应用;文本分类;树桩网络;朴素贝叶斯;TAN
中图分类号:TP391 **文献标识码**:A

文本分类是中文信息处理的一个重要研究领域。其目标是在分析文本内容的基础上,给文本分配一个或多个比较合适的类别,从而提高文本检索、文本存储等应用的处理效率。目前,较为著名的文本分类方法有 Bayes、LLSF、SVM、KNN、决策树等[1]。朴素贝叶斯(native bayes)文本分类模型是一种简单而高效的文本分类模型,但是它的属性独立性假设使其无法表示现实世界属性之间的依赖关系,影响其分类性能。本文讨论了一种改进的贝叶斯文本分类模型"树桩网络(stump network)",并将其与朴素贝叶斯文本分类器、TAN 文本分类器进行实验比较,以测试其分类性能。

1 向量空间模型及基于贝叶斯定理的文本分类

1.1 向量空间模型

在向量空间模型(VSM)中,文档被看作一系列无序词条的集合,对每个词条加上 1 个相应的权值,将文档映射为 1 个特征向量 $V(d)=(t_1,w_1(d);\cdots;t_n,w_n(d))$,其中 t_i 为词条项,为在 d 中的权值。所选用的词条项权值计算方法为 TF-IDF 公式:

$$w_{id}(d)=\frac{tf_{ik}(d)\log(\frac{N}{n_k}+0.01)}{\sqrt{\sum_{k=1}^{n}(f_{ik}(d))^2\times\log(\frac{N}{n_k+0.01})}}。\quad(1)$$

式(1)中,$tf_{ik}(d)$ 表示词条 t_k 在文档 d 中出现的频率,N 表示全部样本文档的总数,n_k 表示包含词 t_k 的文档数。关键词的权值度量着重于那些区分文档类别属性的关键词。每一类的特征向量表示完成后,即完成向量空间模型的建立。待分类文档用同样的方法得到其特征向量。文本转化为向量形式并经特征提取后,便可以进行分类挖掘了。

1.2 基于贝叶斯定理的文本分类

贝叶斯文本分类模型是一种典型的基于统计方法的分类模型[2~4]。贝叶斯定理是贝叶斯理论中最重要的一个公式,是贝叶斯学习方法的理论基础,他将事件的先验概率与后验概率巧妙的联系起来,利用先验信息和样本数据信息确定事件的后验概率。

令 $U=\{W_1,W_2,\cdots,W_N,C\}$ 是离散随机变量的有限集,其中 $W_1,W_2,\cdots,W_N$ 是属性变量,类变量 C 的取值范围为 $\{c_1,c_2,\cdots,c_1\}$,W_i 是属性 W_i 的取值。文档 $d_i=(w_1,w_2,\cdots,w_n)$ 属于类 c_j 的概率,可由贝叶斯定理表示为:

$$p(c_j|w_1,w_2,\cdots;w_n)=\frac{p(w_1,w_2,\cdots;w_n|c_j)\cdot p(c_j)}{p(w_1,w_2,\cdots,w_n)}$$
$$=\alpha\cdot p(c_j)\cdot p(w_1,w_2,\cdots,w_n|c_j)。\quad(2)$$

式(2)中 $w_1,w_2,\cdots,w_n$ 是词汇表中的词,α 是正则化因子,$p(c_j)$ 是类 c_j 的先验概率,$p(w_1,w_2,\cdots,w_n)$ 是 c_j 类的后验概率,先验概率独立于训练样本数据,而后验概率反映了样本数据对类的影响。

依据概率的链规则,式(2)可以表示为:

$$p(c_j|w_1,w_2,\cdots,w_n)=\alpha\cdot p(c_j)\cdot\prod_{i=1}^{n}p(w_i|w_1,w_2,\cdots,w_{i-1},c_j)。\quad(3)$$

收稿日期:2005-10-18
作者简介:王 潇(1980-),女,硕士生,从事数据库研究。e-mail:yuexiaoxue@126.com。

给定训练数据集 $D=\{d_1, d_2, \cdots, d_N\}$,分类任务的目标是对数据集 D 进行分析,确定 1 个映射函数 $f:(W_1, W_2, \cdots W_n) \rightarrow C$,使得对任意未知类别的文本 $d_i=(w_1, w_2, \cdots, w_n)$,可以标以适当的类标 C^*。。

根据贝叶斯最大后验准则,给定某一待分类文本 $d_i=(w_1, w_2, \cdots, w_n)$,贝叶斯分类器选择使后验概率 $p(c_j|\ w_1, w_2, \cdots, w_n)$ 最大的类 C^* 作为该文本的类标签。因此,贝叶斯文本分类模型的关键是如何计算 $p(w_i|\ w_1, w_2, \cdots, w_{i-1}, c_j)$。目前,不同的贝叶斯文本分类模型的区别就在于,他们以不同的方式来求 $p(w_i|\ w_1, w_2, \cdots, w_{i-1}, c_j)$。

1.2.1 基于朴素贝叶斯的文本分类

在朴素贝叶斯文本分类模型中,假定所有的属性变量都是相互类条件独立的,每个节点只与类节点 C 相关联,应此式(3)中 $p(w_i \mid w_1, w_2, \cdots, w_{i-1}, c_j)$ 的简化为 $p(w_i \mid c_j)$。相对于其它分类算法朴素贝叶斯文本分类器的最大特点是不需要搜索,只需简单计算训练例中各个属性值发生的频率数,就可以估计出每个属性的概率估计值,因而朴素贝叶斯分类器的效率特别高。但朴素贝叶斯文本分类器基于 1 个"独立性假定":给定 1 个文本的类标签,文本中每个属性的出现独立于文本中其他属性的出现。在现实中,这种独立性假设经常是不满足的。因此,怎样改进朴素贝叶斯文本分类器,使之在属性独立性假设不满足的情况下依然具有较高的分类精度,就成了一个重要的研究领域。

1.2.2 基于 TAN 的文本分类

TAN 是一种树结构的贝叶斯文本分类模型。在 TAN 结构中,类变量是根,没有父结点,即 $\Pi C=\varnothing$ (ΠC 表示 C 的父节点集),类变量是每个属性变量的父节点,即 $C \in \Pi W_i$ (ΠW_i 表示 W_i 的父节点集,$i=1, 2, \cdots, n$),属性变量 W_i 除了类变量 C 作为其父节点以外,最多有 1 个其它属性变量作为其父节点,即 $|\Pi W_i \leq 2|$。因此,式(3)中的 $p(w_i|\ w_1, w_2, \cdots, w_{i-1}, c_j)$ 或者简化为 $p(w_i|\ c_j)$,或者简化为 $p(w_i|\ w_p, c_j)$。其中 $w_p \in \{w_1, w_2, \cdots, w_{i-1}\}$。

Friedman 等[3]提出了利用条件互信息构造 TAN 分类器的算法。我们以下将介绍的树桩网络也是一种特殊的 TAN 类。

1.2.3 基于"树桩网络"的文本分类

"树桩网络(stump network)"是 Zhang 等提出的一种树状的模型[5]。该模型的灵感来自于现实世界中如果数据集中含有大量的属性,则这些属性倾向于聚集成组。为了反映这种依赖,该算法先将所有属性构造成一个个简单的组,即"树桩",然后再在树桩之间构造连接,该模型适合于含有大量属性的数据集。我们知道,巨大的训练样本和过高的向量维数是文本分类的两大特点,因此我们使用树桩网络来改进贝叶斯文本分类器。

我们首先定义 1 个树结构,然后用贪心搜索策略来发现这个特殊的 TAN 树结构。定义 1(树桩)。令 r 为 1 个属性,N 为属性集。属性 r 是 N 中每个属性的父结点,树的深度为 1,我们把满足以上条件的树结构称为树桩,计为 $T(r, N)$。

定义 2(树桩网络)。任意 2 个"树桩"的交集为空,而树桩之间按照如下方式进行连接:1 个"树桩"的根最多指向另 1 个"树桩"的 1 个叶结点。我们把满足上述条件的一组树桩称为树桩网络。

我们采用启发式的搜索方法[6]来学习"树桩网络"的结构,主要有以下 2 个步骤:

第一步:得到树桩的集合。该集合中的树桩应满足如下条件:如果把这个树桩加入朴素贝叶斯文本分类器能提高分类器在测试集上的分类性能。然后把集合中的树桩按提高分类器的分类性能的大小按降序排列。

第二步:遍历排序后的树桩序列 1 次,看每个树桩的根节点能否指向排序树桩序列中排在它前面的树桩的叶结点,判断的依据是能否提高分类器的分类性能。

其算法的具体实现步骤如下:

1) 读取训练数据集 D 中的数据;

2) 初始化 B 为朴素贝叶斯文本分类器,并估计它的分类性能;

3) 初始化结点集 N(不包括类变量 C)使其包含所有属性,树桩队列为空;

4) 用 N 中的每个节点构造 1 个树桩,设 Ts 为最能提高分类器的分类性能的树桩;

5) 对 Ts 中的每条弧,如果删除这条弧后分类器的分类性能没有减小,就把他从 Ts 中删除;

6) 把 Ts 压入树桩队列;

7) 从 N 中删除 Ts 中的所有节点,如果 N 不为空,转 4;

8) 遍历队列,对队列中的每个树桩,加 1 条从队列中先前树桩的叶节点到该树桩的根节点的弧,如果能提高分类器的分类性能的话就保留,否则就删除。

2　实验结果

我们选用中文自然语言处理开放平台提供的语料库,文本类别为 6 个,依次是计算机、教育、经济、军事、体育、政治,共 1950 篇。取 1500 篇作为训练文档集(training set),余下的 450 篇作为测试集。测试结果见表 1。

从表 1 可以看出,树桩网络文本分类器较朴素贝叶斯和 TAN 在大部分实验数据集上取得了较好的分类性能。

表 1　文本分类测试结果　　%

文本类别	朴素贝叶斯		TAN		树桩网络	
	正确率	召回率	正确率	召回率	正确率	召回率
计算机	93.34	90.12	93.42	90.26	94.89	91.33
教育	92.38	89.11	93.21	90.04	93.41	90.27
经济	90.17	90.25	91.24	91.30	92.03	92.05
军事	95.66	93.29	96.38	93.89	97.06	94.67
体育	99.33	97.42	99.54	97.67	99.47	98.04
政治	90.12	89.11	90.89	90.78	92.28	90.83

3　结论

文本分类面临 2 个明显的问题:训练集中的属性个数很多;属性之间可能存在依赖关系。

朴素贝叶斯文本分类模型特别适合于处理属性个数较多的分类问题,相对于朴素贝叶斯方法,TAN 文本分类方法中增加了表示依赖关系的能力,本文提出的"树桩网络"文本分类方法是一种特殊的 TAN 类,它可以更好的处理文本词语之间存在的依赖关系。本文的实验表明,该方法较前两种有更好的分类性能。

参考文献:

[1] 韩家炜,坎 伯. 数据挖掘:概念与技术[M]. 范 明,孟小峰. 北京:机械工业出版社,2001.

[2] Friedman N. Bayesian network classifier[J]. Machine Learning,1997,29:131-161.

[3] Friedman N, Geiger D, *et al*. Bayesian network classifiers[J]. Machine Learning,1997,29(2-3):131-163

[4] 张　璠. 多种策略改进朴素贝叶斯分类器[J]. 微机发展,2005,15(4):35-39.

[5] Zhang H, Ling C X. An Improved learning Algorithm for Augmented Native Bayes[A]. In:Advances in Artificial Intelligence, LNAL 2903[C]. Berlin Heidelberg:Springer-Verlag,2003.453-456.

[6] 王晓东. 计算计算法与设计[M]. 北京:电子工业出版社,2003.

A Comparative Study on Three Text Classification Algorithms

WANG Xiao, HU Xin

(College of Mathematics and Information Science, Northwest Normal University, Lanzhou, Gansu 730070, China)

Abstract: In this paper, an text classification model based on bayes theorem called stump network is introduced. Stump Network text classifier is compared with naive bayes text classifier and TAN (tree augmented naive bayes) by an experiment. Experimental results show this model has higher classification accuracy in most data sets.

Key words: computer application; text categorization; stump network; naive bayes; TAN