山西财经大学学报(高等教育版)
2002 年 增 刊 Journal of Shanxi University of Finance and Economics SUPPLEMENT 2002
总 第 55 期 HIGHER EDUCATION EDITION Serial No. 55

贝叶斯文本分类方法研究

石洪波 王志海 黄厚宽

[摘 要] 朴素贝叶斯分类器是当前流行的一种文本分类算法,它的简易性使其适合于处理属性个数较多的分类问题; TAN (Tree Augmented Naive Bayes)综合了朴素贝叶斯的简易性以及贝叶斯网表示依赖关系的能力,使其能容纳属性间存在的某种依赖关系。通过实验比较朴素贝叶斯和 TAN,可以发现 TAN 方法具有较好的分类性能。

[关键词] 文本分类; TAN; 朴素贝叶斯; 贝叶斯网; 特征选择

一、引言

文本分类是数据挖掘和机器学习中非常重要的研究领域,文本分类的目标是对新文档标以合适的类标签。由于一个文档可能会与多个文档类相关,所以文档分类问题是多标签分类问题。文本自动分类的过程首先是对训练集中文档的内容进行分析,构造一个分类方案,即分类器。在分类器学习之后,每个类有一个不同的分类方案,可用这些分类方案对新文档分类。

与一般的分类问题相比,文本分类面临两个突出问题:一是属性个数很多,文本分类中的属性指的是文本中的词,在一篇文档中,词的数目是非常大的,少则几百个词,多则成千上万个词;二是属性之间存在某些依赖关系,在某一类文档中,某些词语之间常常存在一定的依赖关系。文本的这些特点,使得文本分类更具挑战性。

朴素贝叶斯分类器是目前公认的一种简单有效的概率分类方法,在某些领域中表现出很好的性能。在朴素贝叶斯分类方法中,有一个"独立性假设":给定一个实例的类标签,实例中每个属性的出现独立于实例中其它属性的出现。这个独立性假设,使朴素贝叶斯方法特别适合处理属性个数很多的任务,而文本分类恰恰就是这种多属性的分类任务。然而,在现实世界中,这个独立性假设明显不成立。于是人们设想能否构造一种模型,避免朴素贝叶斯中不现实的独立性假设,从而提高分类器的性能。这其中的一个关键问题是如何表示属性变量之间可能存在的依赖关系。

Friedman 等人提出了一种新的树状结构模型 TAN (Tree Augmented Naive),其基本思路是放松朴素贝叶斯中的独立性假设条件,借鉴贝叶斯网中表示依赖关系的方法,扩展朴素贝叶斯的结构,使其能容纳属性间存在的依赖关系,但对其表示依赖关系的能力加以限制,使学习该模型成为可能。本文介绍了朴素贝叶斯与 TAN 文本分类方法,并用实验比较了两种文本分类方法,实验表明,该方法具有较高的分类性能。

二、文本表示和特征选择

文本分类是有监督的学习任务,任何文本分类算法在学习之前,都需要将文档以一个合适的表示形式表示出来,使其适应于分类算法。大多数分类方法是基于向量空间模型的,在该模型中,每个文档可看作是词(或词组)的序列,文档中的词称作特征,文档表示为由特征组成的特征矢量。特征矢量中的分量可采用布尔表示(即用 1 表示某特征在文档中出现,用 0 表示某特征在文档中不出现),也可采用频度表示(即表示某特征在文档中出现的次数),还可采用其它一些形式。

考虑提高效率和除去噪音的目的,在文档表示为可用于分类的表示形式之前,需要进行特征选择。特征选择是从每一类文档的所有特征中抽取那些能够反映和区分此类文档与其它类文档的特征项,这是分类问题的关键。文本分类中的特征选择一般是通过大量已知类属的文本,统计出能够反映此类文档的特性。特征选择常采用的处理方法包括:TFIDF、互信息、词频、信息熵等。本文采用的是平均互信息。

三、基于贝叶斯定理的文本分类

(一) 贝叶斯文本分类框架

本文采用布尔表示法描述文档,每个文档由特征矢量表示,V 是词汇表,或者是 1 或者是 0,1 表示词汇表中第 t 个词在文档 i 中出现,0 表示不出现。给定某一文档 d,贝叶斯分类器按下式选择最可能的类 C^*:

$$C^* = \arg\max_{c_j \in C} P(c_j | w_1, w_2, \cdots, w_n) \tag{1}$$

其中 $w_1, w_2, \cdots, w_n$ 是词汇表 V 中的词,根据贝叶斯定理和链规则,上式可写为:

$$\begin{aligned} C^* &= \arg\max_{c_j \in C} \frac{P(w_1, w_2, \cdots, w_n | c_j) \cdot P(c_j)}{P(w_1, w_2, \cdots, w_n)} \\ &= \arg\max_{c_j \in C} P(w_1, w_2, \cdots, w_n | c_j) P(c_j) \\ &= \arg\max_{c_j \in C} P(c_j) \cdot \prod_{i=1}^{n} P(w_i | \prod_{wi}) \end{aligned} \tag{2}$$

其中 $\prod_{wi}$ 是 w_i 的父节点集。

(二) 基于朴素贝叶斯的文本分类

在朴素贝叶斯分类方法中,有一个"独立性假设":给定一个实例的类标签,实例中每个属性的出现独立于实例中其它属性的出现。由于每个属性独立于其它属性,只有一个类节点作为其父节点,所以朴素贝叶斯分类器可表述为:

$$C_{Naive} = \arg\max_{c_j \in C} P(c_j) \cdot \prod_{i=1}^{n} P(w_i) \tag{3}$$

因此,要对一个新文档分类,就是要从训练集中估计出两组概率值:和 $P(c_j)$ 和 $P(w_t | c_j)$。为了提高估计值的可靠性,可以采用 m - 估计或拉普拉斯估计。本文中采用拉普拉斯概率估计,具体的估计公式如下:

$$P(c_j) = \frac{\sum_{i=1}^{|D|} P(c_j | d_i)}{|D|} \qquad j = 1, \cdots, |C| \tag{4}$$

$$P(w_t | c_j) = \frac{1 + \sum_{i=1}^{|D|} B_{it} P(c_j | d_i)}{2 + \sum_{i=1}^{|D|} P(c_j | d_i)} \qquad j = 1, \cdots, |C|, t = 1, \cdots, n \tag{5}$$

其中,D 是训练文档集,$P(c_j | d_i) \in \{0, 1\}$,表示训练文档 d_i 是否属于 c_j 类文档,1 表示属于,0 表示不属于。

(三) 基于 TAN 的文本分类

TAN (Tree Augmented Naive Bayes)是由 Friedman 等人提出的一种树状结构模型,它是朴素贝叶斯分类器的自然扩展。其基本思想是将贝叶斯网的某些表示依赖关系的能力与朴素贝叶斯的简易性相结合,使分类性能增强。

令 $U = \{A_1, A_2, \cdots, A_n, C\}$,其中变量 $A_1, A_2, \cdots, A_n$ 是属性变量,C 是类变量。在 TAN 结构中,类变量是根,没有父结点,即 $\prod_C = \Phi$($\prod_C$ 表示 C 的父结点集),类变量是每个属性变量的父结点,即 $C \in \prod_{A_i}$($\prod_{A_i}$表示A_i的父节点集,i = 1,2,…,n)。属性变量A_i除了类变量 C 作为其父结点外,最多有一个其它属性变量作为其父结点,即$|\prod_{A_i}| \leq 2$。

在模型 TAN 中,由于每个节点最多只能有一个(非类)父节点,所以 TAN 分类器可表述为:

$$C_{TAN} = \arg\max_{c_j \in C} P(c_j) \cdot \prod_{i=1}^{n} P(w_i | \prod_{v_i}) \quad (6)$$

其中 $\prod_{v_i}$有两种形式:(1) $\prod_{v_i} = \{c_k\}$,w_i 没有非类的父节点;(2) $\prod_{v_i} = \{c_k, w_s\}$,$w_i$ 有一个非类的父节点,因此,要对一个新文档分类,就是要从训练集中估计出三组概率值:$P(c_j)$,$P(w_t | c_j, w_s)$以及 $P(w_t | c_j)$。$P(c_j)$和 $P(w_t | c_j)$的估计采用式(4)、(5),$P(w_t | c_j, w_s)$的估计公式如下:

$$P(w_t | c_j, w_s) = \frac{1 + \sum_{i=1}^{|D|} B_{it} B_{is} P(c_j | d_i)}{2 + \sum_{i=1}^{|D|} B_{is} P(c_j | d_i)} \quad j = 1, \cdots, |C|, t, s = 1, \cdots, n \quad (7)$$

四、实验结果

实验数据选取的是国际通用的文本分类标准测试集 Reuters - 21578。该数据集是路透社 1987 年的新闻,经人工汇集和分类,共包含 21 578 篇文档,其中训练集中包含 9 603 篇文档,测试集中包含 3 299 篇文档,另外还有若干篇文档待用,全部文档分为 135 个类别。在训练和测试集中,各个类别的文档中包含的文档数不同。我们选择 8 个类别 ——alum、bop、cocoa、coffee、gas、gnp、oilseed、retail 进行实验。

在训练过程中,首先从每个文档中抽取单词,然后经过 stemming 和移去 stopwords 两步处理,得到每个文档的特征,将所有文档的特征组合在一起得到词汇表。在实验数据集中,有些文档不止属于一个类,可能同时属于多个类,我们简单地将这种文档作为几个文档放入到数据集中。

表 1　朴素贝叶斯与 TAN 的实验结果

	Naive Bayes		TAN	
	precision	recall	precision	recall
Alum	1	0.545	1	0.875
Bop	0.914	0.582	0.855	0.973
Cocoa	0.951	0.773	0.855	0.895
Coffee	0.936	0.793	0.969	0.907
Gas	1	0.571	1	0.981
Gnp	0.866	0.832	0.857	0.882
Oilseed	0.572	0.992	0.953	0.97
retail	0	0	0	0

五、结论

文本分类面临着两个明显的问题:训练集中的属性个数很多,属性之间可能存在依赖关系。朴素贝叶斯特别适合处理属性个数较多的分类问题,相对于朴素贝叶斯方法,TAN 方法中增加了表示依赖关系的能力,可以更好地处理文本词语之间的可能存在的依赖关系。本文的实验表明,TAN 分类器比朴素贝叶斯分类器有更高的分类性能。

[作者单位:山西财经大学　责任编辑:高巍]

(上接第 81 页)

语文教学中的人文精神之所以以这样的方式进行,这是由中国教育的特点决定的。是因为中国缺乏培养“人文精神”的土壤呢?还是“人文意识”在传统的拒斥中失去了根基?笔者认为这个问题不能以是或不是的方式进行回答。早在两千多年前,孔老夫子就已将人文性的教学付诸于实践,他广收门徒,有教无类,对学生因材施教,分门别类,开设“六艺”(礼、乐、射、御、书、数)。这些雏形的人文意识虽然和西方的人文意识还相去甚远,但已说明我国的教育家已开始以人为本的人文探索。这种人本精神在《论语·乡党》中便有所体现:“厩焚,子退朝,曰:‘伤人乎’不问马”。说明中国当时已从敬鬼神转向了重人事,重塑以人为本的价值关怀,这不正是人文性的内质显现吗?到了“五四”时期,中国又在西方思潮的影响下,将人文性的教育纲领提了出来,在 1922 年北洋政府通过的《学校系统改革》中明确提出了人文教育的先进思想:如“追求学生个性发展”,“重视生活教育”等等。那么究竟是何种原因使他们如昙花一样,还没来得及留下绰约的风姿便倏忽而逝,我们不妨从鲁迅的话语中得到一些启示:“尽先输入名词,而并不介绍这名词的意义”“于是各个以意为之”,这或许就是人文精神在语文教育中的曲折迂回坎坷之路。这个特点使得中国的人文性教育在理论上出现了混乱,但同时又打破了囿于外来思想的本身规定,各自按照自己的理解和教育的现实需要来接受并填充人文精神。具体到语文教学中,人文性似乎在中国传统语文教学模式和中国教育状况的双重牵引下,失去了自律,然而几经回落,几经更迭的人文主义教育在现今又被提了出来,如果形象地描述中国人文教育的发展,它就好像是位移中的质点,既有方向,又有大小,但它所表示的只是质点位置的变化,而不是一个新的质点的诞生,所以困扰它发展的本质内源性问题也就被揭示出来:那就是“创新”。

如果一个民族没有“创新”,那么这个民族也就失去了生存与发展的动力,同样,如果语文教育的人文性得不到创新丰富,只是维系着概念间的相互挪用与杂陈状态,语文事业也会在枯竭中走向衰落。当然,在此过程中,语文人文教育所表现出的不成熟与不规范势必会给前进中的语文带来种种困惑与病症,然而,蕴含在其内部的积极因素所表现的对个性的尊重、人格的塑造、本题的回归必然会在反复的实践与砥砺中,向世人交一份满意的答卷:创新,创新,再创新,语文就得救了!

[作者单位:太原成成中学　责任编辑:秦兴俊]