◎数据库、信号与信息处理◎

关联词约束的半监督文本分类方法

韩红旗[1,2],朱东华[1],刘 嵩[1],汪雪锋[1]
HAN Hong-qi[1,2],ZHU Dong-hua[1],LIU Song[1],WANG Xue-feng[1]

1.北京理工大学 管理与经济学院,北京 100081
2.华北水利水电学院 管理与经济学院,郑州 450011
1.School of Management and Economics,Beijing Institute of Technology,Beijing 100081,China
2.School of Management and Economics,North China University of Water Conservancy and Electric Power,Zhengzhou 450011,China
E-mail:bithhq@163.com

HAN Hong-qi,ZHU Dong-hua,LIU Song,et al.Semi-supervised text classification using class associated words.Computer Engineering and Applications,2010,46(4):113-116.

Abstract: A problem is presented to classify unlabeled text documents without training set.Class associated words are the words which represent the subject of classes and provide prior knowledge for training a classifier.A learning algorithm,based on the combination of Expectation-Maximization (EM) and a Naïve Bayes classifier,is introduced to classify documents from fully unlabeled documents using class associated words.In the algorithm,class associated words are used to set classification constraints during learning process to restrict to classify documents into corresponding class labels and improve the classification accuracy. Experiment results show that the technique can solve the problem with much high accuracy,and the classification accuracy with constraints is higher than that without constraints.

Key words: semi-supervised;text classification;class associated words;Expectation-Maximization;Naïve Bayes

摘 要:提出了一种没有训练集情况下实现对未标注类别文本文档进行分类的问题。类关联词是与类主体相关、能反映类主体的单词或短语。利用类关联词提供的先验信息,形成文档分类的先验概率,然后组合利用朴素贝叶斯分类器和EM迭代算法,在半监督学习过程中加入分类约束条件,用类关联词来监督构造一个分类器,实现了对完全未标注类别文档的分类。实验结果证明,此方法能够以较高的准确率实现没有训练集情况下的文本分类问题,在类关联词约束下的分类准确率要高于没有约束情况下的分类准确率。

关键词:半监督;文本分类;类关联词;期望最大化(EM);朴素贝叶斯
DOI:10.3778/j.issn.1002-8331.2010.04.036 **文章编号:**1002-8331(2010)04-0113-04 **文献标识码:**A **中图分类号:**TP181

1 引言

在文本文档分类中经常会碰到这样一类分类问题,给了很多文档,要求将它们分为规定的类别,当然这些文档最初并没有任何规定类别的信息。例如,给定了1 000篇知识产权侵权案例文档,要求将它们按照专利、商标、版权等进行分类。手工分类文档是一项非常痛苦的工作,尤其在文档数量非常多时,例如10 000篇。文本聚类可以在没有任何标注文档训练集的情况下,将文档分为几个簇,但很难保证这些簇与规定的类别一一对应。一种传统的做法是手工将部分文档进行分类标注,将这些文档分为训练集和检验集,然后通过对训练集中已标注文档的学习构造一个分类器,用检验集中的标注文档来验证分类器的准确性,在准确性达到许可条件下,利用这个分类器去分类未标注文档[1-2]。然而,传统的分类器如果要达到较高的准确率,需要大量的标注文档作为训练集,如识别UseNet中新闻组的文档,如果要达到70%的准确率,需要2 000篇标注文档[3],这限制了它在文本分类领域的应用,因为获取大量的标注类别样本常常是困难的、昂贵的和费时的[4-5]。

Nigam等从理论上证明了未标注文档提供的信息有助于提高分类准确性,并介绍了一种方法,结合使用朴素贝叶斯和EM算法,从少量的标注文档和大量的未标注文档中学习分类器,达到了良好的分类准确率[3]。实验结果证明只有40个标注样本,通过学习可以将精度从27%提高到43%[3,6]。但这个方法要求每个类别至少有一个标注文档才能进行学习,有一定数量的标注样本才可以达到较高的准确率,而且选择不同数量的标注样本会对分类器的精度产生较大影响。Blum和Nigam介绍了协同训练的方法,只需非常少的标注文档,便可以构造两个

基金项目:国家软科学计划(the National Soft Science Research Program of China under Grant No.2008GXS3K056)。
作者简介:韩红旗(1971-),男,博士生,讲师,主要研究领域为数据挖掘,信息系统;朱东华(1963-),男,博士生导师,主要研究领域为数据挖掘,科技评价;刘嵩(1975-),男,博士生,主要研究领域为数据挖掘,信息系统。
收稿日期:2009-02-06 **修回日期:**2009-03-26

分类器通过协同学习达到较高的准确率,但协同训练要求能够将数据集按特征分为独立分布的两个分离集[7-8],如果文档特征不能满足这个条件,便没有办法采用这种方法。

从大量文档中去为每个分类类别寻找一定数量的训练样本是一项艰难的工作。如果一个没有任何类别标注的待分类分档集需要按多种形式进行分类,那么采用 Nigam 方法就需要标注多个训练集,手工标注的工作量可能会扩大几倍。既然手工标注是非常艰苦的工作,有没有不用手工标注、利用文档本身的信息来对文档进行分类呢?首先来考查一下手工分类的情况。一般来说,文章都是围绕一定主题开展的,各个主题之间具有相对明显的界限,主题可以通过一些主题词来表示[9],也就是说文档中的主题词对文本分类有很大的帮助。人工分类时一般不必完整地阅读完整篇文章,可以根据文章标题、段落标题和内容中与主题有关的词确定文档所属类别。例如,一篇关于运动的文章中可能会出现运动、篮球、足球、游泳等词汇;一篇关于财经的文章中可能会出现财经、财务、股票、经济、税收等词汇,不同类的主题词一般有很大的不同。把这些与类紧密相关、能反映类主题信息的词汇称为类关联词。显然,类关联词能够提供文档类别的先验知识。

2 文本自动分类的方法

传统的机器学习的文本自动分类技术有监督(Supervised)和无监督(Unsupervised)分类两种[1-2,4-5]。设有一个文档集 $D=\{d_1,d_2,\cdots,d_{|D|}\}$,把每一个文档看作是一个词包(Word bags)[10],用一个词条特征向量表示 $d_i=(w_{d_i1},w_{d_i2},\cdots,w_{d_i|d_i|})$,而忽略文档结构和上下文的语义关系,其中 $w_{d_il}(l=1,2,\cdots,|d_i|)$是词汇集 $V=(w_1,w_2,\cdots,w_{|V|})$中的一个词汇。自动分类的目的就是建立一个分类器 $c=f(d)$,将文档 $d\in D$ 映射到文档类别集 $C=\{c_1,c_2,\cdots,c_{|C|}\}$的一个类别 c 中。监督学习事先要给定一个训练集,训练集中的每个文档都有一个类标号,通过对训练集中的文档分类进行学习构造一个分类器,然后用这个分类器对 D 中的文档进行分类。常见的有监督分类算法有 TFIDF、朴素贝叶斯等。无监督分类一般事先不指定文档类别集 C,而是利用文档特征向量本身的特征,采用某种度量方法,将具有相似特性的数据文档归为簇,使簇内的文档有高的相似度,簇间的文档具有较高的相异度[11]。常见的文本聚类算法有层次凝聚法、平面划分法等[1-2]。

半监督(Semi-Supervised Learning,SSL)文本分类是介于无监督和监督学习之间的学习方法,同时利用标注文档和未标注文档构造一个分类器。标准的半监督学习一般将文档集 D 分为 D_L 和 D_U 两个子集,D_L 中的每个文档都有类标号,而 D_U 中的文档都没有类标号,利用 D_L 构造一个初始分类器,再通过 D_U 的学习改进参数形成一个新的分类器,也有其他一些半监督学习方法,如有约束的半监督学习(SSL with constraints)、直推式半监督学习(Transductive learning)等。常用的半监督文本分类学习的方法有:采用 EM 的生成混合器模型、自训练、协同训练、直推式支持向量机以及基于图表的方式[4-5]。

3 利用类关联词对文档分类的方法

提出的算法基于生成器模型,组合使用朴素贝叶斯与 EM 方法,是 Nigam 算法的一种改进。EM 算法是一类根据最大似然或最大后验估计进行迭代的算法[12],朴素贝叶斯分类器是基于混合器模型的一种著名的文本分类算法。算法从完全未标注文档集出发,利用文档中的类关联词对文档进行划分,形成初始分类器,然后利用类关联词确定的分类约束条件,采用 EM 方法迭代地学习一个分类器,最后给文档标以合适的类别。

3.1 生成器模型

生成器模型是研究较早的一种半监督学习模型[4],它是一种特殊的混合器,混合器的每个分量由不相交的子集构成[3,6]。模型的参数由 θ 定义。在这种模型中,文档 d_i 被看作是一个排序的词串事件,由混合器一个分量按照概率参数 θ 生成。一个文档首先按照概率 $P(c_j|\theta)$选择混合器分量,然后以分布 $P(d_i|c_j)$产生文档 d_i[3]。

$$P(d_i|\theta)=\sum_{j=1}^{|C|}P(c_j|\theta)P(d_i|c_j;\theta)$$

学习分类器是估计生成器的参数,通过对样本集 D 的学习获得模型参数 θ 的最大估计,找到最大的 $\arg\max P(\theta|D)$。基于此模型,Nigam 提出一种组合使用 EM 与朴素贝叶斯的分类算法,建立一个基于多项式混合器模型的分类器。算法需要一个小的标注训练集,首先只根据标注文档训练一个初始分类器,给未标注文档分配类标记,然后利用 EM 算法迭代地利用全部文档训练一个新的分类器。EM 算法迭代地执行 E 步(计算 $P(w_t|c_j;\theta)$和 $P(c_j|\theta)$)和 M 步(计算 $P(c_j|d_i;\theta)$),最终得到一个分类器,按照最大后验概率 $P(c_j|d_i)$对文档进行标注[3,6]。EM 算法是爬山算法,能够保证每一步迭代使结果得到改善,最终得到一个局部最优解[13]。Nigam 的算法基于四个假设:(1)混合器模型;(2)混合器分量与类一对一映射;(3)文档中的单词独立于上下文生成;(4)文档长度对所有类均匀分布。尽管现实中这样的条件并不满足,但实验结果证明,这种方法仍然达到了良好的效果[3,6]。下面提出的算法在 Nigam 提出的模型基础上进行了修改,以适应于从完全未标注类别文档学习分类器。

3.2 类关联词的获取

类关联词的选择是重要的一步。选择区分性能好、代表文档主题的类关联词无疑会对文档集形成很好的划分。宫秀军[9]提到一种利用潜在类别变量进行文档分类的方法,要求用户来提供潜在的类别主题词,每个类标号对应一个潜在类别主题词(类似于类关联词),但实际上一个类关联词可能不足以表达一个类主题,一个类可能需要多个类关联词来说明,另外类关联词的提供不能完全依靠用户,用户提供的类关联词可能缺失、或与实际情况不符。建议结合用户提供和系统推荐来选取合适的类关联词,要根据实际情况选择那些必要的、代表性高的和区分能力强的类关联词。要选取有区分性的类关联词,可采用一种或多种特征选取评价函数,如词频、信息增益、互信息等,对系统词汇进行排名,通过用户交互完成。之所以需要用户提供额外的类关联词,原因在于有些类别的文档可能数量非常少,没有任何一种评价函数会对这类文档中的类别关联词给予较高的评分。

如果文档是 Web 页面,包含的<TITLE>、<H>等标注给出了非常有用的文本分类信息,则类关联词的选择也可以从标注对应的文本中进行选择。这些信息所提供的分类信息也可为分类算法的执行提供一些有价值的限制条件。如第 4 章的实验中,从网上下载的文本是 HTML 格式的,可利用文本标题中包含的唯一类别关联词将文档直接标为相应类别,忽略文档内容中包含的其他的类别关联词。

3.3　利用类关联词对文档集划分

一个类可以包含一个或多个关联词,设最终确定的类关联词用 $Z=\{z_1,z_2,\cdots,z_{|Z|}\}$来表示。根据文档包含的类关联词情况,可以将文档集 D 划分为三组:(1)包含的类关联词同属于一个类;(2)包含的类关联词分属不同的类;(3)不包含类关联词。如果类关联词选择合适,则第一组文档确定无疑地属于某个类,而第二组文档只能属于它的类关联词确定的其中一个类别,不可能属于那些类关联词不在文档中出现的类。把这些条件作为分类限制条件应用在学习算法中,保证包含类关联词的文档不会被划分到无关的类中,能够提高分类的准确性。第 4 章实验中比较了采用和没有采用分类限制条件两种情况下的分类结果。

3.4　利用类关联词分类约束条件学习分类器

在 3.3 节中,利用类关联词将文档集 D 分为不相交的三组。按照划分,第一组中的文档,记为 D_1,以概率 1 属于某个确定的类别;第二组中的文档,记为 D_2,包含多个类的关联词,不能确定属于哪个类,初始可假定等概率地属于每个相应的类;第三组中的文档,记为 D_3,因为不包含任何类关联词,初始可假设它等概率地属于每一个类。用矩阵 $\boldsymbol{B}_{|D|\times|C|}$表示文档和类集之间的初始概率关系,用 C_{d_i}表示文档 d_i 对应的类的集合,则矩阵中的元素 b_{ij} 的初始值可根据文档 d_i 的情况选择公式(1)~(3)计算。

$$\text{当 } d_i\in D_1,b_{ij}=1(c_j\in C_{d_i}),b_{ij}=0(c_j\notin C_{d_i}) \tag{1}$$

$$\text{当 } d_i\in D_2,b_{ij}=\frac{1}{|C_{d_i}|}(c_j\in C_{d_i}),b_{ij}=0(c_j\notin C_{d_i}) \tag{2}$$

$$\text{当 } d_i\in D_3,b_{ij}=\frac{1}{|C|} \tag{3}$$

实际上,矩阵中的元素 b_{ij} 对应 $P(c_j|d_i;\theta)$,初始值是文档所属类别的先验概率。下面就引入修改后的 Nigam 模型,利用 EM 迭代算法通过对文档集的学习估计文档所属类的后验概率。这里仍然采用最大似然估计进行最优判断,似然函数为:

$$l(\theta|D)=\log\frac{P(\theta)}{P(D)}+\sum_{i=1}^{|D|}\sum_{j=1}^{|C|}b_{ij}\log(P(c_j|\theta)P(d_i|c_j;\theta)) \tag{4}$$

EM 算法交替使用 E 步和 M 步来学习分类器参数。第一次迭代前已经形成初始的概率估计参数 $P(c_j|d_i;\theta)$,所以首次迭代时不执行 E 步,只执行 M 步。设已经进行了 k 次迭代,得到估计参数 θ^k,第 k+1 次迭代由以下两个步骤组成:

(1)E 步:基于当前的参数估计 θ^k,考虑到类关联词限定的分类约束条件,对 b_{ij} 进行估计。

$$P(c_j|d_i)=\frac{P(c_j)P(d_i|c_j)}{P(d_i)}=\frac{P(c_j)\prod_{k=1}^{|d_i|}P(w_k|c_j)}{\sum_{r=1}^{|C|}P(c_r)\prod_{k=1}^{|d_i|}P(w_k|c_j)} \tag{5}$$

(2)M 步:基于前一步获得的期望值,最大化当前的参数估计,得 θ^{k+1}。

$$P(w_t|c_j)=\frac{1+\sum_{i=1}^{|D|}N(w_t,d_i)b_{ij}}{|V|+\sum_{s=1}^{|V|}\sum_{i=1}^{|D|}N(w_s,d_i)b_{ij}} \tag{6}$$

$$P(c_j)=\frac{\sum_{i=1}^{|D|}b_{ij}}{|D|} \tag{7}$$

公式(5)使用了下面的独立性假设:(1)文档的词的产生独立于它的内容,即词在文档中出现的位置无先后关系;(2)文档中各个词相对于类别属性是相对独立的。尽管上述假设是非常严格的,但实验证明仍然有良好的分类效果[3,9]。

一般情况下,由类关联词限定的分类约束条件有三类:(1)文档的标题包含且只包含一个类的关联词,文档确定属于该类;(2)当文档只包含一个类的关联词,文档确定属于该类;(3)当文档包含多个类的关联词。在算法的 E 步计算出 $P(c_j|d_i)$ 后,选择公式(7)和(8)计算文档集 D_2 中的文档 d_i 对应的新的 b_{ij} 值。

$$\text{对于 } c_j\notin C_{d_i},b_{ij}=0 \tag{8}$$

$$\text{对于 } c_j\in C_{d_i},b_{ij}=\frac{P(c_j|d_i)}{\sum_{c_k\in C_{d_i}}P(c_k|d_i)} \tag{9}$$

经过迭代后,矩阵 $\boldsymbol{B}$ 中的元素 b_{ij} 实际上是文档所属类别的后验概率 $P(c_j|d_i;\theta)$的估计。可以简单地按最大概率将文档 d_i 标注为相应的类,也可以设定一个阈值,只对后验概率最大的、且高于阈值的文档进行标号,低于阈值的属于待定,然后利用所有已标注文档按照朴素贝叶斯方法学习一个分类器,然后用这个分类器对所有未标注文档进行类别标注。

4　实验结果

通过数据采集程序从国家知识产权局网站的案件报道频道下载了 2005~2008 年全部的知识产权案件(报道数据截止到 2008-12-24),各年份报道的案件数如表 1。

表 1　各年份文档分布

年份	文档数
2005	270
2006	344
2007	327
2008	217
总计	1 158

网上获取知识产权案件文档按年进行了区分,现在要将案例文档分为 10 个类别,以了解中国知识产权案件的基本分布和发展趋势。这 10 个类别是:版权(著作权)、专利权、商标权、集成电路布图设计专有权、植物新品种、制止不正当竞争、厂商名称、原产地名称、货源标注、商业秘密。手工阅读这 1 158 篇文档并进行分类是一件非常麻烦的工作,事前没有提供训练集,没有办法采用监督分类方法,采用传统非监督方法很难保证所分文档恰好落入这 10 个类别中。

在执行文本分类算法前需要进行数据预处理[1-2,14]。首先利用分词工具将文档分词,去掉停用词,总共有 21 449 个词条。根据词频评估函数对特征词条进行排序,通过交互选中类别关联词。根据评估函数利用特征子集选择最终挑选 1 235 个特征词条作为文档的特征向量表示,这里面要包含全部类别关联词。

调用了两次算法,一次没有采用类关联词确定的分类限制条件,一次采用分类限制条件,每次都进行 10 次迭代。分类限制条件要求只能将文档分到自身包含的类关联词对应的类中。表 2 对比了有无分类限制条件下的分类结果。可以看到,利用这种方法,即便有些文档类别只包含很少、甚至一个文档,也可以进行分类,这是 Nigam 方法中所不具备的优点。

为了验证这种方法的准确率,从每年随机选取了 50 个案

表 2 有无限制条件下的分类结果对比

类别	没有限制条件分类	有限制条件分类	完全一致的分类数
版权(著作权)	514	521	503
专利	170	184	170
商标	211	226	207
集成电路	37	20	15
植物新品种	10	7	6
不正当竞争	51	62	44
厂商名称	52	34	27
原产地名称	27	5	3
货源标注	23	17	11
商业秘密	63	82	57
合计	1 158	1 158	1 043

例文档,共选取 200 个文档,然后通过阅读案例文档手工进行了类标号标注,然后对比了有无分类限制条件下的分类准确率,情况见表 3。无类关联词分类限制条件下的分类准确率是 81%,而有分类限制条件下的分类准确率是 83.5%,显然后一种的分类准确率稍高一些。

表 3 有无限制条件下的分类准确率对比

类别	手工标注数	无限制条件			有限制条件		
		数量	一致	正确率/(%)	数量	一致	正确率/(%)
版权(著作权)	110	100	99	90.00	101	100	90.91
专利	25	21	20	80.00	24	22	88.00
商标	45	32	32	71.11	35	33	73.33
集成电路		6			4		
植物新品种	2	4	2	100.00	3	2	100.00
不正当竞争	6	6	2	33.33	9	2	33.33
厂商名称	5	10	2	40.00	6	2	40.00
原产地名称		7					
货源标注	1	3		0	4		0
商业秘密	6	11	5	83.33	14	6	100.00
总计	200	200	162	81.00	200	167	83.50

5 总结

介绍了一种利用类关联词提供的先验信息进行分类的半监督分类方法。分类算法事先不需要提供训练集,从完全未标注文档开始利用用户选择的类关联词进行分类。类关联词在该算法中起着很重要的作用,需要利用用户的先验知识,根据文档中词特征项评价排名和用户提供结合并手工选择来确定。该算法基于生成器模型,组合朴素贝叶斯和 EM 方法实现对未标注文档分类。根据文档内包含的类关联词,将文档分为三类:(1)只包含一个类的关联词;(2)包含两个以上的类的关联词;(3)不包含任何类的关联词,对每一个文档设定初始概率权重,然后通过迭代算法,从全部文档学习一个分类器,最后给每一个文档分配合适的类标签。虽然该算法有较严格的假设条件,实验证明算法仍然达到了较高的分类准确率。在分类过程中可以利用文章标题和文档内容中的类关联词加上分类限制条件,只将文档分配自身类关联词规定的类标签,从而提高了分类准确率。

参考文献:

[1] Han Jia-wei,Kamber M.Data mining:Concepts and techniques[M].2nd ed.New York:Morgan Kaufmann Press,2006.

[2] 史忠植.知识发现[M].北京:清华大学出版社,2000.

[3] Nigam K,McCallum A,Thrun S,et al.Learning to classify text from labeled and unlabeled documents[C]//Proceedings of the Fifteenth National Conference on Artificial Intelligence,1998:792-799.

[4] Zhu Xiao-jin.Semi-supervised learning literature survey.Computer Science,University of Wisconsin-Madison,2008.

[5] Chapelle O,Schölkopf B,Zien A.Semi-supervised learning[M].[S.l.]:The MIT Press,2006.

[6] Nigam K,McCallum A,Thrun S,et al.Text classification from labeled and unlabeled documents using EM[J].Machine Learning,2000,39(2/3):103-134.

[7] Blum A,Mitchell T.Combining labeled and unlabeled data with co-training[C]//Proceedings of the 11th Annual Conference on Computational Learning Theory,1998:92-100.

[8] Nigam K,Ghani R.Analyzing the effectiveness and applicability of co-training[C]//Proc ACM CIKM Int Conf on Information and Knowledge Management,2000:86-93.

[9] 宫秀军,史忠植.基于 Bayes 潜在语义模型的半监督 Web 挖掘[J].软件学报,2002,13(8):1508-1514.

[10] 王珏,周志华,周敖英.机器学习及其应用[M].北京:清华大学出版社,2006.

[11] Jain A K,Murty M N,Flynn P J.Data clustering:A review[J].ACM Computing Surveys(CSUR),1999,31(3):264-323.

[12] Dempster A P,Laird N M,Rubin D B.Maximum likelihood from incomplete data via the EM algorithm[J].Journal of the Royal Statistical Society:Series B,1977,39:1-38.

[13] 张连文,郭海鹏.贝叶斯网络导论[M].北京:科学出版社,2006.

[14] Delen D,Crossland M D.Seeding the survey and analysis of research literature with text mining[J].Expert Systems with Applications,2008,34(3):1707-1720.

(上接 85 页)

参考文献:

[1] Floyd S.IETF RFC3649 Highspped TCP for large congestion windows[S].2003.

[2] 武航星,慕德俊,潘文平,等.网络拥塞控制算法综述[J].计算机科学,2007,34(2):51-53.

[3] 荣亮,王建新.基于控制论的主动队列管理的研究进展[J].小型微型计算机系统,2007,28(11):2039-2040.

[4] 刘明,张鹤颖,窦文华.随机指数标记算法的性能分析与模型控制[J].计算机工程与科学,2005,27(9):66-67.

[5] Stevens R W.TCP/IP illustrated volume1:The protocols[M].范建华,胥光辉,张涛,等译.北京:机械工业出版社,2000:209-243.

[6] Ramkrishnam K,Floyd S,Black D.IETF RFC3168 The addition of Explicit Congestion Notification(ECN) to IP[S].2001:3-48.

[7] Morris R.Scalable TCP congestion control[C]//Proc IEEE INFOCOM 2000,Tel Ariv,Israel.[S.l.]:IEEE Computer Society,2000.

[8] 罗万明,林闯,闫保平.TCP/IP 拥塞控制研究[J].计算机学报,2001,24(1):1-18.