文章编号：2096-1472(2021)-07-28-03
DOI:10.19644/j.cnki.issn2096-1472.2021.07.007

改进的贝叶斯算法在商品分类中的应用研究

邵欣欣

(大连东软信息学院软件工程系，辽宁 大连 116023)
✉Shaoxinxin@neusoft.edu.cn

摘　要：针对采用贝叶斯分类器算法进行商品描述分类时，出现大量混淆性词汇从而无法保证特征间独立的问题，提出了采用停用词优化的贝叶斯分类器算法，通过词频统计和词性筛选的方式，过滤掉大部分混淆性词汇，从而保证特征独立。针对相似类别无法准确区分的问题，提出了子模型训练的解决方案，对易混淆类别单独进行训练并记录训练过程，在测试阶段根据结果判断并使用子模型，从而实现细化区分。实验表明，优化方案确实可行，可以获得97.80%的准确率。

关键词：朴素贝叶斯分类器；停用词；子模型训练；商品分类
中图分类号：TP311　**文献标识码**：A

Research on Application of Improved Bayesian Algorithm in Commodity Classification

SHAO Xinxin

(*Department of Software Engineering*, *Dalian Neusoft University of Information*, *Dalian* 116023, *China*)
✉Shaoxinxin@neusoft.edu.cn

Abstract: A large number of confusing words appear when Naive Bayes classifier is used to classify commodity description. This makes it difficult to ensure independence between features. This paper proposes a Bayesian classifier algorithm optimized by stop words. Most confusing words can be filtered out by using word frequency statistics word part-of-speech screening. Aiming at the problem that similar categories cannot be distinguished, confusing categories are trained separately and the training process is recorded. In test phase, sub-models are judged and then used according to the results, so to realize detailed distinction. Experiments show that the optimized solution is indeed feasible, and can reach an accuracy of 97.80%.

Keywords: Naive Bayes classifier; stop words; sub-model training; commodity classification

1 引言(Introduction)

电子商务的兴起吸引了大量卖家入驻电商平台，此时，传统的商品分类系统就捉襟见肘了。传统分类系统一般依赖于关键字，工作人员根据商品描述为其打上对应的标签，既费时又费力；卖家在提供商品描述时也经常带有主观性，且经常带有新词汇或迷惑性词语，这就导致标签不好给出，并导致推荐算法准确率下降，从而影响下单率。因此，电商平台想要提高下单率、口碑和用户黏性，提供准确的商品分类算法是首要考虑的问题。

国外的文本分类已经到了实用化阶段，并广泛应用于电子邮件分类、垃圾信息拦截、情感分析等方面，其中比较有名的例子是卡内基集团为路透社开发的Construe系统，以及麻省理工大学为美国白宫开发的邮件分类系统。国内对文本分类的研究起步较晚，由于中文无法通过空格对文本进行分词处理，因此，即使是最基本的词频统计算法，也有较高的实现难度，加之很长一段时间内国内都没有一套公开的中文数据集，导致文本分类算法难以实现。经过一段时间的研究，一些机构如清华大学、北京大学都推出了自主建立的汉语语料库，有了这些基础数据，中科院计算所的李晓黎、史忠植等人应用概念推理网进行文本分类，召回率达到94.2%。

基金项目：辽宁省自然科学基金(2019-ZD-0354).

中国科技大学的范众等人在KNN、贝叶斯和文档相似性研究的基础上提出了一个超文本协调分类器，正确率接近80%，特点是适当考虑了HTML文本中结构化信息。复旦大学和富士通研究中心的黄萱菁、吴立德等人研究了独立语种的文本分类，并以词汇和类别的互信息量为评分函数，考虑了单分类和多分类，最好的召回率为88.87%。

目前，一些比较成熟的分类算法已经被广泛用在了文本分类中，主要包括贝叶斯分类、决策树方法、向量机分类、KNN算法[1]等。其中，朴素贝叶斯是一种简单的线性分类器，由于其简单易用的特性和理论基础的支持，在文本分类中得到了广泛的应用。但该算法不足之处在于其假定全部特征之间彼此独立，而商品详情往往会出现不同品类间的类似情况，这一特性导致直接套用该算法并不能生成理想的可用模型。针对上述问题，本文提出了采用停用词和子模型训练的方法对贝叶斯分类算法进行改进，并经实验验证，在商品分类中具有较好的效果。

2　算法描述(Algorithm description)

2.1　朴素贝叶斯分类器算法

朴素贝叶斯分类器算法依赖于条件概率定理，也就是贝叶斯定理，该定理表示如下：

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

其中，$P(A|B)$是指在事件B发生的情况下，A发生的概率。由于已知B发生后A的条件概率，其得自B的取值，因此也称其为A的后验概率。同理，$P(B|A)$称为B的后验概率，也称作似然性。$P(A)$或$P(B)$由于不需要考虑另一条件的关系，因此称作A或B的先验概率[2]。

假设某个个体有n项特征(Feature)，分别为$F_1,\cdots,F_n$；其有i种类别(Category)，分别为$C_1,\cdots,C_n$。贝叶斯分类器的目标是计算出概率最大的那个类别，即：

$$P(C_i|F_1,\cdots,F_n)=\frac{P(F_1,\cdots,F_n|C_i)}{P(F_1,\cdots,F_n)}$$

这个算式的最大值。

由于在同一组数据下，对所有类别来说，$P(F_1,\cdots,F_n)$均相同，可以省略，故问题变为求该公式的最大值：

$$P(F_1,\cdots,F_n|C_i)P(C_i)$$

朴素贝叶斯假设特征之间彼此独立，在这种情况下：

$$P(F_1,\cdots,F_n|C_i)P(C_i)=P(F_1|C_i),\cdots,P(F_n|C_i)P(C_i)$$

等式右侧的每一项均为特征F_n下，类别为C_i的可能性，可以通过数据统计直接从数据中获得，故可以根据此方法找出概率最大的那个类别[3]。

经过测试发现，商品的描述往往存在干扰特征，比如数值数据、商品的寿命、商品的价格等，在不使用停用词的情况下，这些数据用作特征容易混淆不同类别的商品。

2.2　商品分类停用词算法设计

商品分类主要是对文本进行分类，而提高文本分类准确性的主要途径是制定合适的停用词[4]。通过对错误率高的商品描述进行分析，发现影响分类器对商品描述文字信息进行分类的因素主要有以下两种：

自然语言中存在的大量功能词，比如的、地、得、或、但等。这类词汇本身对分类没有帮助，可能会出现在各种类别的商品中，并引起混淆。

商品描述中普遍存在的通用词汇，比如最火、网红、后悔、哭了等。这些词汇与功能词产生的影响是相似的。

针对以上两点，可以大致推理出所需要的停用词方案：

针对功能词：由于功能词比较固定，可以直接选择一套现有符合需求的功能词列表，并添加到系统停用词中。

针对通用词汇：统计全部品类商品的词频，取不同分类中出现最多的相同形容词、副词等修饰性词汇，添加到停用词列表。

针对商品描述中的通用词汇：由于该种词汇具有扎堆出现的特点，因此可以对某段时间内出现某种类扎堆出现的新词汇进行统计。

针对不同子类商品之间存在的混淆词汇：这种情况并不能简单地使用制定停用词列表的方式。

停用词统计算法流程设计如下：

Step 1：读取数据。

Step 2：统计商品详情的词频，并按类别分组。

Step 3：统计每组词汇，找出出现在五个以上组的形容词、副词、量词。

Step 4：统计每组词汇，找出在一段时间内大量出现的相同词汇。

Step 5：将找到的词汇添加到停用词列表。

2.3　商品分类算法设计

经过多组实验，最终得出商品分类算法的全貌。

其中，训练流程为：

Step 1：读取数据，将类别编号，并与商品描述绑定。

Step 2：通过贝叶斯分词器将商品描述进行分词。

Step 3：统计每组词汇，找出出现在五个以上组的修饰性词汇加入停用词列表。

Step 4：统计每组词汇，找出在一段时间内大量出现的相同词汇，加入停用词列表。

Step 5：进行一次标准训练。

Step 6：验证准确性，找出带有易混淆的子分类的主分类。

Step 7：过滤训练数据，收集上一步所找到数据的原始数据，分组进行各一次标准训练。

Step 8：汇总全部模型。

Step 9：保存模型和训练记录。

其中，一次标准训练流程为：

Step 1：过滤分词结果中的停用词。

Step 2：将分词结果利用Tokenizer数值化。

Step 3：使用HashingTF转换器将商品描述的分词结果转换为特征向量。

Step 4：将特征向量转换为朴素贝叶斯分类器算法需要的

格式。

Step 5：代入朴素贝叶斯分类器算法进行训练。

3 算法实施(Algorithm implementation)

3.1 实验环境

本文实现算法的实验环境是Mac OS系统下搭建的Spark环境。Spark版本为2.4.0，Hadoop版本为3.0.3，JDK版本为1.8，Scala版本为2.12.8。

3.2 数据说明

本实验采用的数据为某平台的商品描述数据，包含30万条训练数据以及370万条测试数据，格式为tsv形式。训练数据每行为商品描述、商品类型，测试数据每行为商品描述[5-6]。部分训练数据如表1所示。

表1 数据样例

Tab.1 Data sample

商品描述	商品类别
腾讯QQ黄钻、黄钻3 个月、季卡、自动充值可查时间可续费	本地生活—游戏充值—QQ充值
哈格、宠物狗零食、哈格鸡肉牛奶擦牙卷10 支装、狗狗咬胶洁齿骨、宠物食品	宠物生活—宠物零食—磨牙/洁齿
茶杯紫砂盖杯、景德镇全手工、办公紫砂杯、货到付款、九龙戏珠杯、390 mL	厨具锅具—茶具/咖啡具—茶壶

可以发现，商品类别虽然只有一列，但其实分为三层，且逐层详细：本地充值—游戏充值—QQ充值，依次对应：主分类—子分类—详细分类。

3.3 算法训练

为适当控制训练时间，以下测试将从训练数据中抽取30万条，从测试数据中抽取20万条进行，精确到子分类。

标准朴素贝叶斯分类器模型训练：为保证训练结果的有效性，在该组训练开始前，先进行一次去掉上文中提到的易混淆数据的标准训练，同时下文所涉及的训练数据均为去掉易混淆主分类的数据[7-8]。

分析测试结果发现，在众多商品类型中，女鞋/女装、维修保养/汽车装饰这两组预测的准确率最低，且这两组均为各自同一主分类下的子分类，存在较多重复描述性词汇，还有其他类别也出现了匹配失败的情况，匹配失败结果统计如图1所示。

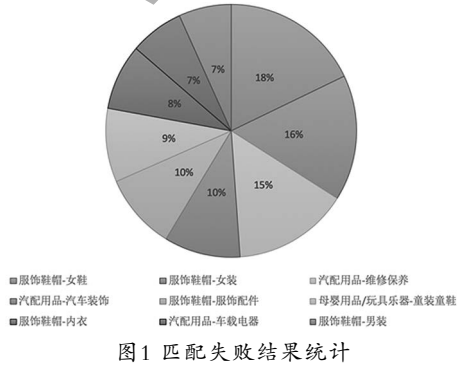

图1 匹配失败结果统计

Fig.1 Matching failure statistics

针对同一主分类中不同子分类间存在重复性词汇的问题，采用子模型训练的方法来解决。

子模型的训练：以服饰鞋帽—女鞋/女装为例，对该组数据进行一次模型训练。分析结果为针对两种子类别进行训练的准确率远高于对全品类进行分类时的准确率。当训练数据只包含子品类时，朴素贝叶斯分类器能够更好地分析出两者的不同。由此可初步认为，对子分类单独成组进行训练以区分易混淆的商品类别是可行的。该子模型将会保存并在稍后使用。

项目选用以下两种标准停用词列表进行测试：

(1)使用哈尔滨工业大学提供的汉语语言停用词列表对商品详情的分词结果进行过滤。

(2)使用百度提供的百度搜索停用词列表对商品详情的分词结果进行过滤。

百度搜索停用词表是针对搜索引擎定制的停用词表，由于商品分类的目标是实现商品搜索功能，故选用该停用词表进行测试[9]。

根据实验结果，发现除上文中所提到的易混淆子分类外，其他分类的错误率有一定程度的下降，但同样未达到预期效果。经过以上两次测试，可以初步得出结论：汉语语言停用词列表并不是通用的，停用词列表跨领域使用效果并不理想，只是使用通用的解决方案并不能很好地过滤商品分类中的无用词汇。

由于使用通用的停用词解决方案无法满足需求，因此将对整个商品分类数据构建自己的停用词。收集两类词语作为本系统的停用词列表，一是同一段时间扎堆出现的新词汇；二是不同分类中出现最多的相同形容词、副词等修饰性词汇。收集到的停用词如下：最新、热销、热门、活动等。

从数据集中随机抽取30万条，70%作为训练集，30%作为测试集，进行实验，采用多种方法进行分类，各个方法的准确率和用时情况如表2所示。

表2 算法准确率和用时情况比较

Tab.2 Comparison of accuracy and time

序号	算法	准确率	用时
1	标准朴素贝叶斯分类算法	87.70%	2 小时24 分钟
2	使用哈尔滨工业大学提供的汉语语言停用词列表对商品详情的分词结果进行过滤	91.80%	2 小时30 分钟
3	使用百度提供的百度搜索停用词列表对商品详情的分词结果进行过滤	92.79%	2 小时30 分钟
4	采用本文训练的停用词表	95.42%	2 小时57 分钟
5	使用该算法进行三次迭代	97.80%	4 小时30 分钟

分析实验结果，发现准确率已经达到预期，不过由于进行了数次子训练，导致时间较长。实际上在应用时会直接使

(下转第27页)