基于朴素贝叶斯的新闻分类改进

孙子杰
（中国人民大学附属中学，北京，100080）

摘要：步入了大数据时代，可接收到的信息越来越多。面对海量的信息，无论是新闻的阅读者还会新闻网站的工作人员，往往都面对这一个问题——文本分类。人工分类耗时耗力，且工作效率随时间增加而下降，这些缺点无不将这件难题推向计算机来解决。本文选择朴素贝叶斯算法，以多篇新闻为实验训练样本进行实验，结合数据预处理、汉语分词等过程得出一个较完整的数学模型，并对数据平滑技术提出改进，为今后学者的研究提供一个可参考的方案。

关键词：朴素贝叶斯；文本分类；汉语分词

0 引言

随着时代的发展，信息爆炸一词早已不再陌生，互联网上丰富的信息给人们的生活带来诸多方便，例如降低了学习资料获取的成本等；与此同时，信息爆炸的负面性也渐渐的影响着我们的生活，铺天盖地的广告、新闻、信息远远超出我们的接受范围，其中充斥着大量的无用甚至虚假信息。从海量的信息中抽取对自己有用的信息这件事占用了现代人大量的时间，因此，如何高效的进行信息分类成为亟待解决的问题。

常见的信息呈现方式有视频、图片、文本等，本文针对文本信息进行研究。文本分类属于自然语言处理的范畴，自然语言的处理是现阶段研究的热门难题，而汉语结构的复杂性和几千年来汉语中沉淀的人文历史，让汉语文本的分类难上加难。

文本分类作为很常见的热门难题，自然积累了大量的技术实现方法。由于素材和数据集可采集性较高，训练集标记难度不大，现有的文本分类水平也在不断跟进，简单来说，文本分来的过程是通过将文章分段、分词，对词语的词性、词义等进行判断，以小见大达到对整个文章内容进行分类。具体分为基于规则的方法和基于统计的方法。

基于规则的方法是基于研究人员（例如语言学家）对语言的规律进行总结，形成规则形态的知识库，但是由于语言的复杂性，导致很难选取一个规则覆盖所有的语言现象，社交媒体不规范的语言使用习惯也使得基于规则的方法效率较低；基于统计的方法也叫基于机器学习的方法、经验主义方法，是一种机器从语言样本中自动学习的方法，其利用统计技术或机器学习技术，利用语料库训练语言模型。传统机器学习方法往往结构简单，执行简洁，原理明确，对硬件要求略低，在文本分类问题上更受欢迎。目前，学界对自然语言处理有了多种方法，如 N 元模型，KNN、隐马尔科夫模型、神经网络深度学习等等模型。

朴素贝叶斯算法对小规模的数据表现较好，适合多分类任务及增量式训练，因此本文从知名的新闻网站上获取新闻文本数据集，运用基于朴素贝叶斯的模型进行实验，尝试将近 200 篇文章分为 9 个分类，基于朴素贝叶斯模型，探索了一些对数据平滑技术的改进办法，通过一些精度调整，使模型的结果与实际情况更加吻合，为文本的分类提供更加优质的预测方法，提高分类的准确度。

多次试验的结果对比证明了数据预处理的重要性，因此，本文针对数据预处理做出较详细阐述，并根据汉语文本类数据的独有性质，选用流行的可视化库 Matplotlib 将文本数据可视化，清洗，选择，归约等预处理。在实验过程中，采用了交叉验证等方法避免过拟合，最终得到一个较好的预测结果。经过检验，本文提供的方法，可以在一定程度上提高对文本进行多分类的稳定性和准确性。

1 数据处理

■ 1.1 介绍数据与数据预处理

数据（data）是指对客观事物观察并进行记录的结果，是对客观事物的性质、状态及相互关系等进行逻辑归纳的物理符号。显然，数据是一个抽象概念，其具有规模和属性。数据规模是指数据的多少，如今大火的“大数据”即指规模极大，非常复杂的数据；数据属性是指数据所具有的性质，数据具有的性质越多，即属性越多，或称维度越大，维度过大的数据中常包含一些无关属性，此时便需要进行数据降维处理以达到筛选的目的。与其他属性一样，数据也有描述单位，生活中所说的手机容量，网速等等都应用到了数据的单位方面的内容。

杂乱的数据需要经过加工后才能成为信息，那自然就需要在正式计算前进行数据预处理。现实世界中的数据通常比较杂乱，无法直接进行带入算法计算，为了提高数据分析的最终效果，需要对原始数据进行处理。瑕疵数据通常产生于输入时的遗漏 、系统默认值、人工疏忽、噪声、设备 / 系统故障等等原因。在大量的数据中，往往还存在着数据缺失和数据冗余。数据缺失是指某些可能的相关因素被忽略从而导致分析结果与实际出现偏差。数据冗余是由于一些不相关的因素混杂其中，成为干扰因素，对分析可能会造成不必要的

影响，需要剔除。而大数据的作用就是尽可能的搜集齐所有的影响因素，分析数据越精炼越好。

■ 1.2 数据预处理的方式

预处理形式分为数据清理，数据集成与变换，数据归约和离散化及概念分层，除此之外，还存在其他可能需要数据预处理的情况，例如数据的压缩存储，数据形式的转换和数据内容的筛选和梳理等。本文详细仅介绍数据清理，数据集成，数据变换与数据归约。

1.2.1 数据清理

现实世界的原始数据一般是不完整、有错乱的。数据清理试图填充遗漏的值，识别并消除噪音，并更改数据中的不一致为一致。

处理遗漏值常采用的措施有：忽略此个元组、人工补充遗漏值、使用全局常量、平均值等填充某个遗漏值等。

噪音是指测量变量的随机错误或偏差。去除噪音需要数据平滑技术，包括分箱，聚类，计算机和人工检查结合，回归等。分箱是指存储的值被分布到一些“桶”或箱中，通过考察周围的值来平滑箱中存储数据的值。聚类是将类似的值组织成群或“聚类”，落在聚类集合之外的值被视为噪声。计算机和人工检查结合即计算机根据可能的错误模式进行预搜索，人工对错误模式进行检验。回归则可以通过让数据适合一个函数（如回归函数）来平滑数据，如线性回归，找出拟合两个变量的直线，使得一个变量能够预测另一个。

1.2.2 数据集成

数据集成是指将多个数据源中的数据集合，放在一个一致的数据存储中。

数据集成主要根据数据的相关性进行判断，数据相关性包括强正相关、弱正相关、强负相关、弱负相关、非线性相关和不相关。其中，强正相关是指共同增加或减少，且变化明显，说明 x 是 y 的主要影响因素；弱正相关是指共同增加或减少，但变化不明显，说明 x 是 y 的影响因素，但不是唯一因素；强负相关、弱负相关与正相关相反；非线性相关是指 x、y 没有明显线性相关关系，但有某种非线性相关关系，x 仍是 y 的影响因素，不相关即二者完全无关。

1.2.3 数据变换

数据变换分为规范化，数据泛化，属性构造，平滑，聚集等。常见的规范化例如最小 - 最大规范化是指将原始数据 A 经过线性变换，映射到区间 [new_minA, new_maxA]，这种映射存在一个问题，若存在离群点，可能影响规范化，若在规范化后添加新的数据，当新数据落在原数据的区间 [minA, maxA] 之外，将导致“越界”错误。

数据泛化是一个过程，它将数据集从较低的概念层抽象到较高的概念层。泛化的规则为：存在大量不同值，且属性值无法概念分层则删除；存在大量不同值，属性值可以概念分层，则将属性值概念分层；存在少量不同值则保留；不存在不同值则删除。

1.2.4 数据归约

大数据环境下数据量太大，直接进行复杂的数据分析和挖掘效率太低，因此需要更强大的计算能力，更高效的挖掘方法并且减少数据量但并不损失数据特征。研究证明，归约后的数据集上的挖掘结果与原结果几乎相同。

归约策略包括（1）数据立方体聚集：对数据立方体做聚集操作 （2）属性子集选择：检测并删除不相关、弱相关或冗余的属性和维，使得数据类的概率分布尽可能地接近使用所有属性得到的原分布（3）维度归约：数据仅有部分的维与挖掘目标相关，去掉不相关的维。属性维的选择算法包括向前选择，向后删除，二者结合等（4）数值归约： 通过数值特征代替其他数据。

■ 1.3 分析本文所用新闻文本数据

本项目数据来自新闻中的各类文章，分为财经、IT、健康、体育、旅游、教育、招聘、文化、军事九个种类，本文尝试将近 200 篇文章分为以上 9 个分类。

对于从网站上爬虫得到的文本数据，通常带有 html 标签，需要去除标签，本文使用 Python 的 BeautifulSoup 包进行去标签操作。文本分类使用的是词特征，所以本文选择 jieba 中文分词工具对文章进行分词，jieba 分词采用动态规划查找最大概率路径，找出基于词频的最大切分组合，对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用 Viterbi 算法，分词结果如图 1 所示。

图中结果可以明显看到有一些对文章实际意义无关的词，其不具有任何类别表征能力，因此需要停用词去除，比如“的”、“了”等常见连词。去掉停用词后进行词频统计，词频统计是进行特征提取、特征权值计算的基础，根据词频

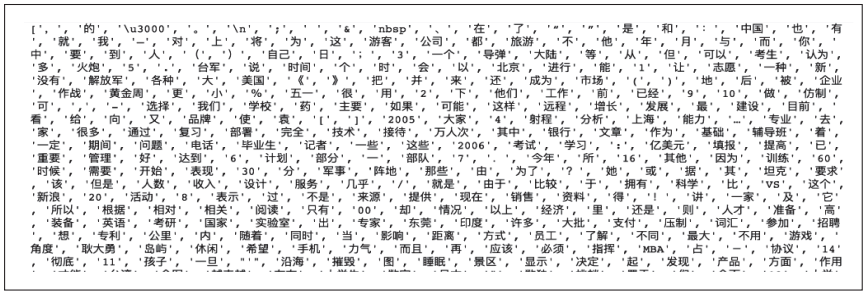

图 1

统计去掉最高频的前 100 个词，此时，数据已经可以为模型所用。

2 模型引入

■ 2.1 朴素贝叶斯

朴素贝叶斯算法（Naive Bayes）是有监督的学习算法，解决的是分类问题，如客户是否流失、是否值得投资、信用等级评定等多分类问题。该算法的优点在于简单易懂、学习效率高、在某些领域的分类问题中能够与决策树、神经网络相媲美。

但由于该算法以自变量之间的独立（条件特征独立）性和连续变量的正态性假设为前提，就会导致算法精度在某种程度上受影响。

朴素贝叶斯模型基于贝叶斯决策理论，用 p1(x,y) 表示数据点 (x,y) 属于类别 1(图 2 中圆点表示的类别) 的概率，用 p2(x,y) 表示数据点 (x,y) 属于类别 2(图 2 中三角形表示的类别) 的概率，那么对于一个新数据点 (x,y)，可以用下面的规则来判断它的类别：如果 p1(x,y)>p2(x,y)，那么类别为 1；如果 p1(x,y)<p2(x,y)，那么类别为 2。也就是说，我们会选择高概率对应的类别。这就是贝叶斯决策理论的核心思想，即选择具有最高概率的决策。

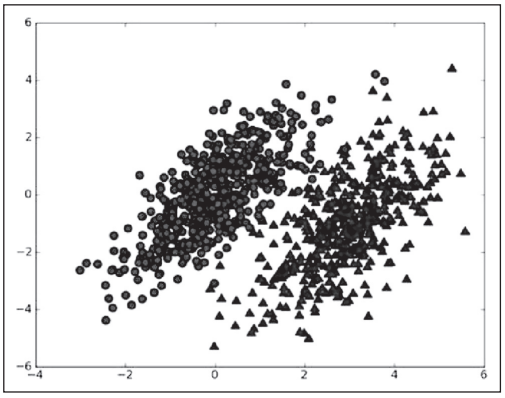

图 2

2.1.1 条件概率与全概率

条件概率是指事件 A 在另外一个事件 B 已经发生条件下的发生概率。条件概率表示为：P（A|B），读作“在 B 的条件下 A 的概率”。若只有两个事件 A，B，那么

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

全概率是指，如果 A 和 A’ 构成样本空间的一个划分，那么事件 B 的概率，就等于 A 和 A’ 的概率分别乘以 B 对这两个事件的条件概率之和，即

$$\square\square \quad = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A')P(A')}$$

2.1.2 贝叶斯推断

在学界，通常把 P(A) 称为“先验概率”（Prior probability），即在 B 事件发生之前，对 A 事件概率的一个判断，P(A|B) 称为“后验概率”（Posterior probability），即在 B 事件发生之后，对 A 事件概率的重新评估。P(B|A)/P(B) 称为“可能性函数”（Likelyhood），作为调整因子，使得预估概率更接近真实概率。所以，条件概率可以理解成下面的式子：

后验概率 = 先验概率 × 调节因子

贝叶斯推断的含义即先预估一个“先验概率”，然后加入实验结果，看这个实验到底是增强还是削弱了“先验概率”，由此得到更接近事实的“后验概率”。

2.1.3 朴素贝叶斯模型

贝叶斯和朴素贝叶斯的概念是不同的，区别就在于“朴素”二字，朴素贝叶斯对条件个概率分布做了条件独立性的假设，贝叶斯分类器的基本方法：在统计资料的基础上，依据某些特征，计算各个类别的概率，从而实现分类。

■ 2.2 问题与改进

在实验中，对常见的三个问题进行改进，分别为平滑问题，下溢出问题，和准确率提升问题。

平滑问题源于一些需检测词在词表中未出现导致后验概率为 0，这显然是不合理的，常见的解决方法为拉普拉斯平滑（又称加一平滑），即规定出现次数比真实次数多一次，使得未出现的词组概率不再是 0，而是大于 0 的较小的值，但是，对所有没出现过的词组都增加同样的频次，并不合理，对于量级较小的数据，规定出现次数比真实次数多一次次数过多，因此，本文选择效果更好的 Add-k 平滑，即不再是加 1 次而是视情况而言加 k 次，实验结果证明，Add-k 结果优于 Add-1，将分类准确率由 73% 提升至 81%。

下溢出问题是由于太多很小的数相乘造成的，在程序中，在相应小数位置进行四舍五入，计算结果可能就变成 0，为了解决这个问题，对乘积结果取自然对数，通过求对数可以较好的避免下溢出或者浮点数舍入导致的错误。

准确率的提升选择 Bagging 策略，Bagging 策略来源于 bootstrap aggregation：从样本集（假设样本集 N 个数据点）中重采样选出 N_b 个样本（有放回的采样，样本数据点个数仍然不变为 N），在所有样本上，对这 n 个样本建立分类器，重复以上两步 m 次，获得 m 个分类器，最后根据这 m 个分类器的投票结果，决定数据属于哪一类。

3 结语

本实验选用的朴素贝叶斯模型优点较多，朴素贝叶斯模

（下转第 52 页）

R2(config-router)#router-id 2.2.2.2
R2(config-router)#network 12.1.1.0 0.0.0.255 area 2
R2(config-router)#network 23.1.1.0 0.0.0.255 area 1
R2(config-router)#network 2.2.2.2 0.0.0.0 area 1
R2(config-router)#area 1 virtual-link 3.3.3.3
R2(config-router)#exit
R3 上的配置：
R3(config)#router ospf 1
R3(config-router)#router-id 3.3.3.3
R3(config-router)#network 23.1.1.0 0.0.0.255 area 1
R3(config-router)#network 34.1.1.0 0.0.0.255 area 0
R3(config-router)#area 1 virtual-link 2.2.2.2
R3(config-router)#exit
R4 上的配置：
R4(config)#router ospf 1
R4(config-router)#router-id 4.4.4.4
R4(config-router)#network 34.1.1.0 0.0.0.255 area 0
R4(config-router)#network 4.4.4.4 0.0.0.0 area 0
R4(config-router)#exit

当配置完成以后，Lo1 和 Lo2 就可以正常通信了。本实验完成了 OSPF 多区域的配置和虚链路的配置，其中涉及到很多 OSPF 的具体原理，值得深入研究。OSPF 的路由分为 3 种类型，分别是域内路由、域间路由和外部路由，其中外部路由又分为一类外部路由和二类外部路由。它们之间的优先级排序为域内路由、域间路由、一类外部路由和二类外部路由。

以下是 OSPF 协议中 Hello 报文的结构，通过 Wireshark 进行分析可以看出 Hello 报文中每个字段的值，但由于篇幅有限，在此不作赘述。总之，通过对协议的分析和解释，我们可以很好的理解 OSPF 协议的工作原理和运行机制。

```
Open Shortest Path First
▾ OSPF Header
    Version: 2
    Message Type: Hello Packet (1)
    Packet Length: 48
    Source OSPF Router: 1.1.1.1
    Area ID: 0.0.0.2
    Checksum: 0xcc8d [correct]
    Auth Type: Null (0)
    Auth Data (none): 0000000000000000
▾ OSPF Hello Packet
    Network Mask: 255.255.255.0
    Hello Interval [sec]: 10
  ▸ Options: 0x12 ((L) LLS Data block, (E) External Routing)
    Router Priority: 1
    Router Dead Interval [sec]: 40
    Designated Router: 12.1.1.1
    Backup Designated Router: 12.1.1.2
    Active Neighbor: 2.2.2.2
▾ OSPF LLS Data Block
    Checksum: 0xfff6
    LLS Data Length: 12 bytes
  ▸ Extended options TLV
```

图 2　Hello 报文的结构

3 结束语

OSPF 是目前使用最广的一种内部网关路由协议，被广泛应用于企业的核心网络当中，作为网络运维人员必须要很好的掌握。本文介绍了 OSPF 协议仅仅是基本的原理和配置方法，部分高级原理并未涉及，望读者能够先夯实基础，多做实验，通过不断的练习最终达到掌握 OSPF 协议的目的。

参考文献

* [1] 郭佳 . 多区域 OSPF 路由协议实验的设计与实现 [J]. 科技创新与应用 .2017.
* [2] 章丞 . 基于 GNS3 的金融行业 OSPF 路由流量设计与实现 [J]. 福建电脑 .2018.
* [3] 黄沈炜 .OSPF 路由技术原理及网络设计探讨 [J]. 中国新通信 .2017
* [4] 宋尚 .OSPF 区域间环路问题 [J]. 信息与电脑 (理论版).2017

（上接第 39 页）

型为生成式模型，通过计算概率来进行分类，可以用来处理多分类问题，其对小规模的数据表现很好，适合多分类任务，适合增量式训练，算法的逻辑也比较简单。

当然，在一些情况下它也存在着不足，朴素贝叶斯推断的常见缺点例如，对输入数据的表达形式很敏感，由于朴素贝叶斯的“朴素”特点，所以会带来一些准确率上的损失和需要计算先验概率，分类决策存在错误率等。

参考文献

* [1] 李晓菲 . 数据预处理算法的研究与应用 [M]. 西南交通大学，2006.
* [2] 徐光美 . 用平滑方法改进多关系朴素贝叶斯分类 [J]. 计算机工程与应用 , 2017 , 53 (5) :69-72.
* [3] 谢斌 . 朴素贝叶斯分类在数据挖掘中的应用 [J]. 兰州文理学院学报 (自然科学版), 2007 , 21 (4) :79-82.
* [4] 苏金树 . 基于机器学习的文本分类技术研究进展 [J]. 软件学报，2006.
* [5] 刘晓霞 . 文本分类中互信息特征选择方法的研究 [J]. 计算机工程与应用 , 2010 , 46 (34) :123-125.
* [6] 楼巍 . 面向大数据的高维数据挖掘技术研究 [D]. 上海大学 , 2013.