朴素贝叶斯及其改进算法在垃圾邮件过滤中的应用

詹鹏伟,谢小姣
(广州大学 数学与信息科学学院,广东 广州 510006)

摘 要 朴素贝叶斯模型在文本分类领域应用广泛,但因为算法本身的缺陷,分类性能有待提高。文章在传统的朴素贝叶斯模型的基础上,利用对数处理解决了算术下溢问题,使用拉普拉斯平滑解决了因训练集过小出现的零概率问题,并采用了系数加权的方法改善了朴素贝叶斯因假设所有条件都是独立的而导致的性能问题,进一步根据垃圾邮件过滤必须要有的查准率高的特点提出了阈值限定条件,最终训练的出的模型分类效果较传统的朴素贝叶斯模型有所提高,对垃圾邮件过滤模型的设计有一定的指导作用。

关键词 朴素贝叶斯;系数加权;阈值限定
中图分类号:TP393.0 文献标志码:A 文章编号:2095-2945(2018)20-0157-03

Abstract: Naive Bayesian model is widely used in the field of text classification, but the classification performance needs to be improved because of the defects of the algorithm itself. Based on the traditional naive Bayesian model, the problem of arithmetic underflow is solved by logarithmic processing, and the zero probability problem due to the small training set is solved by Laplacian Smoothing. The method of coefficient weighting is used to improve the performance of naive Bayes, which is caused by the assumption that all the conditions are independent. Furthermore, the threshold limit condition is proposed according to the characteristic of the high precision rate necessary for spam filtering. The classification effect of the final training model is improved compared with the traditional naive Bayes model, which can guide the design of spam filtering model.

Keywords: naive Bayes; coefficient weighting; threshold qualification

引言

随着互联网的发展,电子邮件的使用也越来越普及,但是电子邮件的安全性与可靠性却还有待提高。各种钓鱼邮件、垃圾邮件、骚扰邮件极大的影响了我们的日常生活,根我国网络不良与垃圾信息举报受理中心的统计,超过半数的用户会因为垃圾邮件而浪费时间、浪费资源,将近一半的用户则有可能因为垃圾邮件而遭受经济损失。可见,设计一个性能良好的垃圾邮件过滤器将有很重要的现实意义。

在文本分类领域,朴素贝叶斯模型有着重要的应用。得益于其简单有效,能够实现增量式计算且对缺失数据敏感度较低的优点,特别适合构建垃圾邮件的过滤模型,但由于其所假设的特征项之间的独立性,将会对最终的结果的准确率产生一定的影响,且未经平滑处理的贝叶斯模型在小数据集上容易出现较大的误差。本文将采用一种有阈值限制的条件的基于系数加权的改进贝叶斯模型,改善了传统贝叶斯模型的分类性能,实现对垃圾邮件的准确过滤。

1 朴素贝叶斯模型及相关原理

1.1 贝叶斯定理

贝叶斯定理是用于描述的是两个不同的事件 A、B 间,A 为条件 B 发生的概率与 B 为条件 A 发生的概率之间的关系。贝叶斯公式可表示为:

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

其中 P 为事件发生的概率。利用贝叶斯定理来构造的决策方法是在所有相关概率都已知的情况下,考虑如何基于这些概率和可能的期望损失来选择最优分类的方法。现假设有 N 种可能的类别$\{c_1,c_2,\cdots,c_N\}$,且存在样本 $x\in\{x_1,x_2,\cdots,x_N\}$,需要将样本 x 分为相应的类别,则可以定义基于后验概率 $P(c_i|x)$ 将某一样本 x 分类为 c_i 所产生的期望损失:

$$R(c_i|x)=\sum_{j=1}^{N}\lambda_{ij}P(c_j|x)$$

其中 λ_{ij} 表示将真实类别为 c_j 分类为 c_i 所产生的损失。利用贝叶斯定理来分类的目标是:寻找能够最小化全局风险的准则 h,h 应为:

$$h(x)=\text{argmin}R(c|x),c\in\{c_1,c_2,\cdots,c_N\}$$

即在每个样本 x 上都选择能使得期望损失 R 最小的类别 c,此时为所得到的贝叶斯分类器的性能上限。

在利用贝叶斯定理来最小化期望损失相当于是利用有限的训练样本尽可能准确的估计后验概率 $P(c|x)$ 的过程,而基于贝叶斯定理,后验概率 $P(c|x)$ 可表示为:

$$P(c|x)=\frac{P(c)P(x|c)}{P(x)}$$

其中 $P(c)$、$P(x)$ 称为先验概率,$P(x|c)$ 为样本 x 相对于类别 c 的条件概率,则求后验概率 $P(c|x)$ 的过程就转化为了在给定的数据集中估计先验概率 $P(c)$ 和条件概率 $P(x|c)$ 的问题,这相比于后验概率 $P(c|x)$ 来说更容易实现。

1.2 极大似然估计

对于上一节中所叙述的贝叶斯概率模型,训练过程就

是对模型的参数进行估计的过程，而根据频率主义学派的思想，极大似然估计可以用于解决参数估计的问题。

假设 D_c 为数据集合 D 中类别为 c 的样本的集合，则需要估计的数据集 D_c 对于参数 θ_c 的条件概率，即似然为：

$$P(D_c|\theta) = \prod_{x \in D_c} P(x|\theta)$$

极大似然估计就是寻找一个能使数据出现的可能性最大的 θ_c。

1.3 算术下溢问题

在使用计算机进行极大似然估计时，有大量的概率值乘法计算，有可能出现算术下溢问题，导致所计算出来的后验概率具有不确定性，使得参数估计与预期相差甚远，最终的分类性能也会大大下降。为了解决算术下溢问题，可将极大似然估计的目标进行对数化处理，则所用到的对数极大似然估计为：

$$\log(\theta) = \log(P(D_c|\theta)) = \sum_{x \in D_c} \log(P(x|\theta))$$

$$\theta_c = \arg\min \log(\theta)$$

1.4 朴素贝叶斯模型

在利用贝叶斯定理来估计后验概率 $P(c|x)$ 时，由于 $P(x|c)$ 是所有属性上的联合概率，难以估计。朴素贝叶斯则假定各属性之间都是条件独立的，将后验概率的计算方式转变为了：

$$P(c|x) = \frac{P(c)}{P(x)} \prod_{i=1}^{d} P(x_i|c)$$

其中 d 为属性的数量。在此假设下，所需要寻找的准则 h 转变为了：

$$h(x) = \arg\min P(c) \prod_{i=1}^{d} P(x_i|c)$$

以上即为朴素贝叶斯分类器的表达式。

1.5 拉普拉斯平滑

因为在训练贝叶斯模型的时候存在大量的连乘，在训练集不够全面的情况下，可能因为某个特征从未在训练集中出现过，导致该特征对预测概率的贡献值为 0，进一步导致最终计算出来的概率为 0，分类结果会产生较大的误差，称这种情况为零概率问题。特别是在将邮件文本作为向量的情况下，虽然构造的词表很大，但是文字的组合更多，极有可能出现未在词表中出现的词语，导致最终的预测结果为不准确。

为了解决这种情况，需采用拉普拉斯平滑（Laplacian smoothing），即在计算先验概率与似然的时候为每个特征出现的次数加上一个很小的数，这样对最终的结果影响不大，且在数据集足够大的时候，产生的概率变化可以忽略不计，但是却很好的解决了零概率问题。在拉普拉斯平滑下，所计算的先验概率 $P(c)$ 与 $P(x_i|c)$ 变为了：

$$P(c) = \frac{|D_c| + 1}{|D| + N}$$

$$P(x_i|c) = \frac{|D_{c,x_i}| + 1}{|D_c| + N_i}$$

其中 D 表示训练集合，D_c 表示训练集中类别为 c 的样本组成的集合，D_{c,x_i} 表示 D_c 中出现的取值为 x_i 的样本集合，N 表示 D 中出现的可能的类别数，N_i 表示第 i 个属性可能的类别数，所加的常数 1 即为解决零概率问题而选择的一个很小的常数。

2 朴素贝叶斯模型的改进

2.1 系数加权

朴素贝叶斯模型假设所有的条件都是相互独立的，也就是所有条件对于最终结果的贡献程度都是一样的，但是在现实中是过于理想的。为了避免这一局限性，对最终结果贡献程度较大的特征不应该与贡献程度较小的特征采取同样的处理方式，在这一思想下，本文采用了一种基于系数加权的改进贝叶斯模型，来为模型中的不同特征赋予部不同的权重，从而突出贡献率大的特征的作用。如在垃圾邮件的识别中，如果发现邮件中出现了类似于“优惠”、“购买”、“理财”等词语，则这封邮件有很大概率是一封垃圾邮件，就算这封邮件中的其他词语是在垃圾邮件中比较少见的，也必须对这封邮件提起警觉。在这一思想下，朴素贝叶斯模型中的准则 h 转变为了：

$$h(x) = \arg\min P(c) \prod_{i=1}^{d} w_i P(x_i|c)$$

其中 w_i 为每个属性的权重。

为了使用基于系数加权的朴素贝叶斯模型，首先需要先确定权值，其方法为：对出现的每个属性直接用朴素贝叶斯分类器进行分类，分类得到的正确率即为权值 w_i。

2.2 阈值限定条件

一个实用性强的垃圾邮件过滤器必须在不将重要邮件误归为垃圾邮件的前提下，将垃圾邮件尽可能的过滤出来，即分类的结果必须有极高的查准率（Precision）。若是用户的重要邮件被垃圾邮件过滤器给过滤掉了，则可能会耽误用户的正事，给用户带来极大的损失。相比之下，若是有少数的垃圾邮件未被成功过滤，虽然对用户的使用体验有影响，但是不至于给用户带来直接的损失，在遇到这类很有可能是垃圾邮件但是确定性不够高的邮件时，可以专门作出提醒，但是不进行过滤。

为了提高垃圾邮件分类的查准率，在本文中采取的垃圾邮件的改进朴素贝叶斯分类器还新增了一个阈值条件：当一封邮件被归为垃圾邮件的概率至少为该邮件被归为正常邮件概率的 1.3 倍时，才有足够的把握将一封邮件归为垃圾邮件过滤掉，否则只进行提醒，而不进行过滤，既：

$$\begin{cases} P(是|x) \geq 1.3 \times P(否|x), & 过滤 \\ P(否|x) \leq P(是|x) < 1.3 \times P(否|x), & 做出提醒（不过滤） \\ P(是|x) < P(否|x), & 不过滤 \end{cases}$$

3 分类结果与分析

为了能够将本文所用的分类器与传统的朴素贝叶斯分类器的性能进行比较，首先定义参数查准率 P 与查全率 R：

$$P = \frac{确实为垃圾邮件的数目}{识别出的垃圾邮件数量}$$

$$R=\frac{\text{确实为垃圾邮件的数目}}{\text{样本中的垃圾邮件数量}}$$

本文采用的实验数据集为 CCERT 提供的邮件数据集，从中随机选取 1000 封正常邮件和 1000 封垃圾邮件，并经过纯文本化处理，去除 html 标签以及其他不相关项。首先利用 Python 接口的结巴分词库完成词向量表的构建，然后将邮件转换为对应的向量表示，再分别利用传统的朴素贝叶斯模型与本文提出的带有阈值限制的系数加权的贝叶斯模型进行训练与预测。实验环境为 macOS High-Sierra、Intel Core i5、1.8GHz 主频、4GB 内存 Python3.6、结巴分词 v0.39。最终所得到的结果如表 1 所示。

表 1 传统贝叶斯与改进贝叶斯的分类情况

	朴素贝叶斯模型	改进的贝叶斯模型
P	92.18%	94.10%
R	90.56%	95.81%
F1	91.36%	94.94%

可见最终的分类结果不管是 P 值还是 R 值都有一定的提高，其中 P 值提高了将近 2%，R 值提高了超过 5%，综合两者的综合评价 F1-Measure 值从 91.36%上升到了 94.94%，可见，改进的贝叶斯模型对垃圾邮件的分类效果较好。在查准率提高了的情况下，用户的正常邮件被误分类的几率下降了，该垃圾邮件分类器的实用性有所提高，减少了因为用户的正常邮件被过滤而造成损失的几率。但是因为在运算中新增的加权计算，训练模型所用到的时间较传统的贝叶斯模型有所增加，当邮件数量较大时耗费的时间较久，速度有待提高。

4 结束语

垃圾邮件过滤器必须在不将正常邮件分为垃圾邮件的基础上尽可能的过滤垃圾邮件，即必须要有较高的查准率才具有实用价值。本文在传统的朴素贝叶斯模型的基础上，解决了算术下溢问题与应训练集较小而导致的零概率问题，并采取了系数加权的方法来改善朴素贝叶斯模型因假设所有特征都是相互独立所导致的分类性能问题，进一步提出阈值限定条件，来改善模型的查准率，一定程度上提高了分类的性能。虽然改进之后的模型在训练时间上有所增加，但是分类性能有所提高，对垃圾邮件过滤的应用有一定的指导作用。

参考文献：

[1]王青松，魏如玉.基于短语的贝叶斯中文垃圾邮件过滤方法[J].计算机科学，2016，43（04）：256-259+269.
[2]杨杉，何跃，颜锦江.基于贝叶斯的反垃圾邮件技术探讨[J].网络安全技术与应用，2007（08）：54-56.
[3]刘牛.基于属性加权的朴素贝叶斯分类算法改进[J].网络安全技术与应用，2011（06）：72-74.
[4]郑炜，沈文，张英鹏.基于改进朴素贝叶斯算法的垃圾邮件过滤器的研究[J].西北工业大学学报，2010，28（04）：622-627.
[5]秦锋，任诗流，程泽凯，等.基于属性加权的朴素贝叶斯分类算法[J].计算机工程与应用，2008（06）：107-109.

（上接 156 页）

保护，解决了常规保护整定困难、配合困难的问题，提高了保护的选择性和灵敏性。

⑷ 区域后备保护采用动态自适应技术，实现保护范围的动态自适应和定值免整定，适应电网分布式能源的接入、环网运行等各种电网网运行方式，提高电网的智能化程度。

⑸ 区域电网智能自愈、安全稳定控制的应用，实现了在区域内对负荷转供、低频减载、联机切负荷、备自投等功能的自由组合。

4 结束语

本文介绍了智能变电站区域化保护的应用。采用区域保护控制方案，变电站只需配置就地采集和执行单元，站内设备数量大幅度减少，显著减少变电站建筑面积。简化了网络结构，提高了变电站自动化系统的可靠性。区域化保护系统的实施将改善现有继电保护性能和安全稳定控制水平，提高系统运行的安全性和可靠性。

参考文献：

[1]李俊刚，张爱民，彭华厦，等.区域层次化保护系统研究与设计[J].电力系统保护与控制，2014，42（11）：34-40.
[2]陈宇.区域备自投发展及在云南电网推广应用前景[J].云南电力技术，2017，45（5）：40-42.
[3]金震，董凯达，张军，等.基于实时信息的区域备自投研究与实现[J].中国电力，2016，49（12）：76-80.
[4]周伊琳，孙建伟，陈炯聪，等.区域网络备自投及其测试关键技术[J].电力系统自动化，2012，36（23）：109-113.
[5]宋保泉，张延鹏，张武样，等.智能变电站区域差动保护技术研究[J].东北电力技术，2013，11：12-14.
[6]于静，信鹏飞.浅谈智能变电站继电保护技术的优化[J].科技创新与应用，2017（11）：212.
[7]韩海山，王莉.关于智能变电站继电保护技术优化措施探讨[J].科技创新与应用，2017（03）：198.