融合朴素贝叶斯与决策树的用户评论分类算法

贾晓帆,何利力

(浙江理工大学 信息学院,浙江 杭州 310018)

摘 要: 为了实现对用户评论的商业研究价值提取,解决互联网产品后续优化和增进服务问题,提出一种融合朴素贝叶斯与决策树的改进算法,处理文本中的噪声,避免零概率和属性值缺失的问题,从而提高分类准确率。该算法首先对用户评论数据作预处理,然后运用概率优化后的朴素贝叶斯处理空缺属性值,最后用决策树从积极和消极角度将数据进行分类。对微信公众号用户评论数据集进行实验,结果表明改进后的算法准确率达80.27%,比传统方法提高0.5%。

关键词: 用户评论分类;决策树算法;朴素贝叶斯

DOI: 10.11907/rjdk.202244 **开放科学(资源服务)标识码(OSID):**

中图分类号: TP312 **文献标识码:** A **文章编号:** 1672-7800(2021)007-0001-05

User Comment Classification Algorithm Based on Naive Bayes and Decision Tree

JIA Xiao-fan, HE Li-li

(*School of Informatics and Electronics, Zhejiang Sci-Tech University, Hangzhou* 310018, *China*)

Abstract: In order to extract the business research value of user reviews and solve the problems of subsequent optimization and service improvement of Internet products, an improved algorithm combining naive Bayes and decision tree is proposed to deal with the noise in text and avoid the problems of zero probability and missing attribute values, so as to improve the classification accuracy. Firstly, the algorithm preprocesses the user comment data, then uses the probability optimized naive Bayes to deal with the missing attribute values, and finally uses the decision tree to classify the data from the positive and negative perspectives. The experimental results show that the algorithm achieves a 80.27% accuracy rate of 0.5% compared with the traditional method through experiments on WeChat official account user reviews dataset.

Key Words: user review classification; decision tree algorithm; Naive Bayes

0 引言

随着互联网的飞速发展,为了满足用户需求,出现了网页、软件、手机应用等互联网产品,还包括建立在各类平台上而开发出的产品,如微信小程序、公众号等。用户在互联网中发表对产品的评价这一举动让用户从单一的信息接受者转变为互联网中文本信息的发布者,文本信息量呈指数级增长,仅仅由人工进行分析提取几乎不大可能,如何有效管理并充分利用这些信息值得思考。

朴素贝叶斯是机器学习的一个常用分类模型,模型本身是建立在贝叶斯定理和特征条件独立假设上的,有着坚实的数学基础,用概率统计知识对样本数据集进行分类。1990年,Kononenko等[1]证明了朴素贝叶斯的有效性。朴素贝叶斯的优势在于能够很快地在训练集中建立起贝叶斯模型,但是在有些实际应用中分类效果却不尽如人意。因为在用贝叶斯分类的前提下,必须假设属性独立,即属性之间没有关系,当该假设不成立时,就会影响贝叶斯分类效果。为了解决该问题,学者们放松属性之间相互独立的条件假设,提出了贝叶斯网络分类器[2],其基本思想是考虑全部或者部分属性之间的关联性,以此满足朴素贝叶斯模型相互独立的条件假设。尽管这种思想能提高分类性

收稿日期: 2020-11-03

基金项目: 国家重点研发计划项目(2018YFB1700702)

作者简介: 贾晓帆(1996-),女,浙江理工大学信息学院硕士研究生,研究方向为智能计算与数据挖掘;何利力(1966-),男,博士,浙江理工大学信息学院教授、博士生导师,研究方向为制造业信息化、企业智能。

能，但是在训练中需要测算所有属性之间的相关性，导致算法复杂度剧增。1996年，Sahami[3]提出K-依赖贝叶斯分类器，有效提升了分类性能；1997年，Friedman等[4]提出了一种树扩展的朴素贝叶斯分类器，简称TAN模型，它在测算属性之间相关性的基础上，构建树形结构图；1999年，Nurnberger等[5]提出了基于神经模糊的朴素贝叶斯分类器，简称BAN模型，它是TAN模型的升级版，相比TAN模型，允许特征属性之间形成有向结构图；2004年，Wang等[6]提出基于自适应的Boosting与朴素贝叶斯结合方法，可以有效地缓解噪声提高分类性能；2008年，徐光美等[7]通过基于互信息计算各属性与各类别之间的相关性，选择不相关的特征值代入朴素贝叶斯模型中；2011年，Zheng等[8]通过删除一部分相关性强的特征属性，将处理后的特征属性应用于朴素贝叶斯分类模型；2014年，杜选[9]提出一种基于加权补集的朴素贝叶斯文本分类模型，这种模型可避免在训练集不平衡时，可能导致分类性能低的缺陷。

本文在研究朴素贝叶斯算法的基础上提出一种与决策树相融合的算法，使用本文算法可以有效填补数据集中的缺失属性值。实验结果证明，本文算法的分类效果比传统朴素贝叶斯分类效果更好。

1 理论基础

1.1 Laplace方法

Laplace是最古老的平滑技术之一，所谓平滑技术是指为了产生更理想的概率以调整最大似然估计的技术[10-11]。计算公式如下：

$$PLap\left(w_1\cdots w_n\right)=\frac{C\left(w_1\cdots w_2\right)}{N+T} \tag{1}$$

其中，N为训练实例总数量；T为训练集实例种类数。在Laplace估计中，先验概率P(Y)被定义如下：

$$p\left(Y=y_j\right)=\frac{n_c+k}{N+n*k} \tag{2}$$

其中，n_c是满足$Y=\left\{y_j\right\}$的实例个数，N是训练集个数，n是类的个数，并且k=1。

1.2 朴素贝叶斯

朴素贝叶斯分类(NBC)是一种假设特征与特征之间相互独立的算法，它基于贝叶斯定理，算法逻辑稳定且简单，朴素贝叶斯的健壮性较好，其分类功能在数据展现出不同特点时，差别不大，也即在不同类型的数据集中不会表现出太大差异[12]。因此，当数据集属性之间的关系相对独立时，朴素贝叶斯分类算法会有较好的效用。

(1)贝叶斯定理。根据条件概率可知，事件A在事件B已发生的条件下发生概率为：

$$P\left(A|B\right)=\frac{P\left(A\cap B\right)}{P\left(B\right)} \tag{3}$$

同样地，在事件A已发生条件下事件B发生的概率为：

$$P\left(B|A\right)=\frac{P\left(A\cap B\right)}{P\left(A\right)} \tag{4}$$

结合两个方程式可以得到：

$$P\left(A|B\right)P(B)=P(A\cap B)=P(B|A)P(A)$$

上式两边同除以P(A)，若P(A)为非零，可以得到贝叶斯定理：

$$P\left(B|A\right)=\frac{P(A|B)P(B)}{P(A)} \tag{5}$$

(2)朴素贝叶斯定理。设有样本数据集$D=\left\{d_1,d_2,\cdots,d_n\right\}$，对应样本数据的特征属性集为$X=\left\{x_1,x_2,\cdots,x_d\right\}$，有类别集合$Y=\left\{y_1,y_2,\cdots,y_m\right\}$，即$D$可以分为$y_m$类别。其中$x_1,x_2,\cdots,x_d$相互独立且随机，则$Y$的先验概率$P=P\left(Y\right)$，$Y$的后验概率$P=P\left(Y|X\right)$，由朴素贝叶斯算法可得后验概率为：

$$P\left(Y|X\right)=\frac{P(Y)P(X|Y)}{P(X)} \tag{6}$$

朴素贝叶斯基于各特征之间相互独立，在给定类别为y的情况下，式(6)可以进一步表示为：

$$P\left(X|Y=y\right)=\prod_{i=1}^{d}P\left(x_i|Y=y\right) \tag{7}$$

由式(7)可计算出后验概率为：

$$P\left(Y|X\right)=\frac{P(Y)\prod_{i=1}^{d}P\left(x_i|Y=y\right)}{P(X)} \tag{8}$$

由于$P(X)$的大小固定不变，因此在比较后验概率时，只比较式(8)的分子部分即可。因此，可以得到一个样本数据属于类别y_i的朴素贝叶斯计算公式如下：

$$P\left(y_i|x_1,x_2,\cdots,x_d\right)=\frac{P\left(y_i\right)\prod_{j=1}^{d}P\left(x_j|y_i\right)}{\prod_{j=1}^{d}P\left(x_j\right)} \tag{9}$$

1.3 决策树

决策树是一种包含根节点、内部节点、叶节点3种节点的树型结构。当文本分类算法为决策树时，它由树的内部节点逐一标注所构成，叶节点表示对应的类别标签，与叶节点相连的分支上标注着其对应的权重，树的叶子节点表示文本分类目标，当从根开始遍历查询最终到达某一个叶子节点，这样就完成了一次文本分类[13]，树的高度就是时间复杂度，它是一个自顶向下、分而治之的总过程。决策树的准确率会因重复的属性而受一定影响，因而用决策树进行分类是要对数据进行特征选择。决策树在开始阶段会浪费时间，但只要模型建立起来，运用阶段非常快。

决策树算法是一种无监督分类方法，决策树的生成主要分为节点分裂和阈值确定，节点分裂指当一个节点所代表的属性无法判断时，则选择将一节点分为多个子节点，而选择适当的阈值可以使分类错误率最小。决策树以树的层次规则为特征，叶子节点为分类目标，通过遍历根节

点到叶子节点完成一次分类操作。决策树分类算法与其他决策支持工具相比较起来易于理解和解释。然而，通过有限的数据集无法训练出可靠的类标签，并且对特征空间计算成本非常高，这对决策树而言是一种限制[13]。决策树算法作为一种常见分类算法，有很多变种，包括ID3、C4.5、C5.0、CART等。其中，最常用的、最经典的是C4.5算法。

2　改进思想

2.1　朴素贝叶斯概率优化

朴素贝叶斯的概率估计会在训练样本不足时出现零概率的问题，这会导致求出的类标签的后验概率值为零，使用连乘计算文本出现概率时也是零，传统的朴素贝叶斯在这种情况下无法进行分类，这一直以来都是朴素贝叶斯的难题。为了解决该问题，法国的数学家拉普拉斯（Laplacian）首次提出一种加法平滑方法，对每个词的计数加1，该加法平滑也叫拉普拉斯平滑。该方法基于一定的数学理论基础，在数据集较大的情况下，每一个词出现的次数加1之后对概率估计结果的影响可完全忽略不计，但是却可以有效地避免出现零概率问题。针对本实验所采取的朴素贝叶斯模型，采用词频估算每一个特征，其经过拉普拉斯平滑方法处理后的表达式为：

$$P\left(x_i|c_j\right)=\frac{\sum_{k=1}^{|D|}N\left(x_i,D_k\right)+1}{\sum_{i=1}^{|V|}\sum_{k=1}^{|D|}N\left(x_i,D_k\right)+|V|} \tag{10}$$

2.2　朴素贝叶斯与决策树融合

为了提高贝叶斯分类准确性，在提出的贝叶斯与其他方法相结合的算法中，最为代表的是Kohavi[14]提出的NBTree算法。对于大型数据分类，决策树和朴素贝叶斯相比，前者在维度较大或者属性之间的依赖关系明显优于后者，后者的分类结果准确性优于前者。而NBTree算法刚好结合贝叶斯和决策树各自的优点，提升了算法效率。

C4.5算法的优点是产生的分类规则易于理解，准确率较高，C4.5经过树生成和树剪枝建立决策树。在计算每个属性的信息增益率（Information GainRatio）后，选出信息增益率最高的属性对给定集合测试属性，再采用递归算法根据测试属性建立分支，初步得到决策树。

在用决策树对测试样本数据进行分类时可能会有某些属性值缺失，传统的决策树算法在面对这些缺失的属性值一般会采用抛弃缺失属性值或者重新给定一个在训练样本中该属性常见的值[15]。而C4.5算法会采用概率分布填充法处理缺失属性值，具体执行过程：首先为某个未知属性每个可能的值赋予一个概率，再计算某节点上属性不同值的出现频率，这些概率可以被再次估计[16]。C4.5算法相关计算公式如下所示[17]：

（1）期望信息（也称信息熵）。设S是S_i个数据样本的集合，假定类标号属性有m个不同值，定义m个不同类$T_i\left(i=1,\cdots,m\right)$。设$S_i$是类$T_i$中的样本数。对一个给定的样本分类所需的期望值为：

$$Info\left(S\right)=-\sum_{i=1}^{m}\frac{S_i}{S}\log_2\frac{S_i}{S} \tag{11}$$

（2）信息增益。由属性A划分成子集的信息量为：

$$E\left(A\right)=Info_A\left(S\right)=\sum_{j=1}^{v}\frac{S_i}{S}Info(S_{ij}) \tag{12}$$

信息增益为原来的信息需求与新的需求之间的差。即：

$$Gain\left(A\right)=Info\left(S\right)-E(A) \tag{13}$$

（3）信息增益率。C4.5算法中引入了信息增益率，属性A的信息增益率计算公式为：

$$GainRatio\left(A\right)=\frac{Gain(A)}{SplitE(A)} \tag{14}$$

$$SplitE\left(A\right)=-\sum_{i=1}^{k}\frac{\left|S_i\right|}{|S|}\log_2\frac{\left|S_i\right|}{|S|} \tag{15}$$

在决策树的剪枝阶段，C4.5算法采用后剪枝技术形成决策树模型，根据建立好的模型生成一系列IF-THEN规则，实现对训练集的分类。

C4.5算法虽然在处理噪声方面有很强的能力，但是在训练集缺失属性值很高的状态下，使用C4.5算法构建的决策树模型会变复杂并出现更多的结点数，最终分类准确率也会下降[18]。鉴于朴素贝叶斯分类具有坚实的理论基础、较小的出错率，本文提出一种基于朴素贝叶斯定理的方法[19]，以处理空缺属性值。

假定一个样本训练集$D=\left\{d_1,d_2,\cdots,d_n\right\}$，每一个训练实例描述为$d_i=\left\{d_{i1},d_{i2},\cdots,d_{ih}\right\}$，对应样本训练的特征属性集$A=\left\{A_1,A_2,\cdots,A_n\right\}$，每个属性$A_i$的属性值是$\left\{A_{i1},A_{i2},\cdots,A_{ih}\right\}$。训练集包含的类别集合$C=\left\{C_1,C_2,\cdots,C_m\right\}$，即$D$可以分为$C_m$类别。与训练集D有关系的决策树有如下特点：①内部节点由A_i表示；②子节点属性与父节点属性用枝干相连；③叶节点由C_i表示。树被建立起来后对每个测试实例进行分类，实例d_i的结果就是一个类别。基本步骤为：一是用训练集构造一个决策树；二是将概率优化后的朴素贝叶斯运用到测试集D中。

对于训练集D，首先运用决策树分类器对每个d_i进行分类。如果训练集中没有空缺属性，则将数据压入D1集合，如果有，则按空缺个数压入D2集合，数量越少越排前，到所有数据都检测完则结束，最后形成了D1和D2两个集合，其中D1放入的是没有空缺属性值的数据集合，D2放入的是包含空缺属性值的数据集合；读取D2中的数据，用概率优化后的朴素贝叶斯处理空缺属性后放入D1中，递归直到D2集合中数据为0。处理完训练集D中的空缺属性值后，再用决策树对更新后的训练集进行分类。

3 实验分析

3.1 实验数据

本文实验选取的是某互联网营销平台微信公众号中活动的91 120条中文用户评论，通过考察用户评论，将数据分为积极和消极两类数据，并选用准确率和召回率评价分类效果。部分数据如表1所示。

Table 1 Data of some user comments

表1 部分用户评论数据

id	用户评论数据
1	生活少不了阳光，美丽橙色的阳光更添生活的多姿多彩
2	想象着生命中的色彩，大概就是橙色，活力四射，向着梦想出发~~~
3	越忙碌越精彩的活出美好生活的瞬间原来橙色是如此多娇美妙生活中缺一不可

3.2 数据预处理与文本特征抽取

数据预处理包括3个部分，即：文本正则化、切分成词和去掉停用词。运用TF-IDF方法抽取数据特征，如表2所示。

Table 2 Data preprocessing and text feature extraction

表2 数据预处理与文本特征抽取

id	用户评论数据	经过数据预处理后
1	生活少不了阳光，美丽橙色的阳光更添生活的多姿多彩	生活/少不了/阳光/美丽/橙色/阳光/更添/生活/多姿多彩
2	想象着生命中的色彩，大概就是橙色，活力四射，向着梦想出发~~~	想象/生命/中/色彩/大概/橙色/活力/四射/梦想/出发
3	越忙碌越精彩的活出美好生活的瞬间原来橙色是如此多娇美妙生活中缺一不可	越/忙碌/越/精彩/活出/美好生活/瞬间/橙色/多娇/美妙/生活/中/缺一/不可

3.3 朴素贝叶斯与决策树分类融合

采用融合朴素贝叶斯和决策树的算法对用户评论进行分类，最终正确地将用户评论分为积极和消极两类，如表3所示。

Table 3 Classification results

表3 分类结果

类别	结果
积极	活动搞得好，我要大力支持。每天都在期盼着
	平和，是一种淡定！品质是一种态度！楼外楼，一吸难忘！爱你如初！
消极	这发的是什么，叫我们商户关注，就发些这种没用的，麻烦来点实用的，例如门面如何经营之类的文章，我们也会去看，不然关注你们有什么用
	不好喝，苦苦的

本文实验结果采用准确率和召回率两个指标，计算公式如下：

准确率(Accuracy)：(TP + TN) / (TP + FP + TN + FN)

召回率(Recall)：TP / (TP + FN)

其中，TP表示正确的标记为正，FP错误的标记为正，FN错误的标记为负，TN正确的标记为负[20]，如表4所示。

Table 4 Parameter meaning

表4 参数含义

真实情况	预测结果	
	正例	负例
正例	TP	FN
反例	FP	TN

最终计算得到的分类准确率和召回率如表5所示。

Table 5 Analysis of experimental results

表5 实验结果分析

项目	准确率(%)	召回率(%)
融合朴素贝叶斯与决策树用户评论分类	80.27	78.35
朴素贝叶斯分类	79.75	78.12

由表5可知，为了对比本文中提出的算法是否可行有效，通过对比朴素贝叶斯对用户评论的分类，本文算法准确率高出0.5个百分点，召回率高出0.2个百分点。由此可见，在用户评论文本分类中，融合朴素贝叶斯和决策树用户评论分类效果取得了不错的结果。

4 结语

为了实现对用户评论的商业研究价值提取，解决互联网产品后续优化和增进服务问题，本文提出了一种融合朴素贝叶斯与决策树的用户评论分类算法。该研究首先对文本正则化、切分成词并去掉停用词，再融合朴素贝叶斯和决策树算法，并将其应用于微信公众号互联网营销用户评论分类中，最终可以正确地将用户评论分为积极和消极两类。其中，积极的用户评论可以作为后续互联网营销活动优化，提升用户体验的重要参考依据，消极的用户评论可以增进自己的服务。对分类结果的分析表明，改进后的算法解决了朴素贝叶斯的零概率问题和决策树因属性值缺失率高导致的分类准确度下降问题，提高了分类准确率。由于数据集不足以及中文语义复杂，会造成评论分类出现相反的情况，后续要在语义情感特征准确性提取上作进一步研究。

参考文献：

[1] KONONENKO I. Comparison of inductive and Naive Bayesian learning approaches to automatic knowledge acquisition[J]. Current Trends in Knowledge Adquisition，1990，1：190-197.

[2] LI S H，ZHANG J. Review of Bayesian networks structure learning [J]. Application Research of Computers，2015，32(3)：641-646.
李硕豪，张军. 贝叶斯网络结构学习综述[J]. 计算机应用研究，2015，32(3)：641-646.

[3] SAHAMI M. Learning limited dependence Bayesian classifiers[C]. The 2th International Conference on Knowledge Discovery and Data Minning(KDD)，1996：335-338.

[4] FRIEDMAN N, GEIGER D, GOLDSZMIDT M. Bayesian network classifiers[J]. Machine Learning,1997,29(2-3):131-163.

[5] NUMBERGER A, BORGELT C, KLOSE A. Improving naive Bayes classifiers using neuro-fuzzy learning[C]. 6th International Conference on Neural Information Processing, 1999:154-159.

[6] WANG L M, YUAN S M, LI L. Boosting naive Bayes by active learning[C]. Proceedings of 2004 International Conference on Machine Learning and Cybernetics,2004:1383-1386.

[7] XU G M, YANG B R, QIN Y Q,et al. Multi-relational Naive Bayesian classifier based on mutual information[J]. Chinese Journal of Engineering,2008,30(8):963-966.
徐光美,杨炳儒,秦奕青,等.基于互信息的多关系朴素贝叶斯分类器[J].北京科技大学学报,2008,30(8):963-966.

[8] ZHENG F, WEBB G I. Tree augmented naive Bayes[M]. Boston:Encyclopcdia of Machine Leaning,2011.

[9] DU X. Rsearch on weighted complement-based Naive Bayes text classification algorithm[J]. Computer Applications and Software,2014,31(9):253-255.
杜选.基于加权补集的朴素贝叶斯文本分类算法研究[J].计算机应用与软件,2014,31(9):253-255.

[10] LI D M. Research on hybird classification based on Navie Bayes and Decision Tree[D]. Dalian: Dalian Maritime University,2016.
李冬梅.朴素贝叶斯与决策树混合分类方法的研究[D].大连:大连海事大学,2016.

[11] CESTNIK B. Estimating probabilities: a crucial task in machine learning[C]. Proceedings of ECAI, 1990, 90:147-149.

[12] MA G. Improvement and application of Naïve Bayes algorithm[D]. Hefei: Anhui University,2018.
马刚.朴素贝叶斯算法的改进与应用[D].合肥:安徽大学,2018.

[13] SONG X M. Research on chinese information classification based on improved Bayesian algorithms [D]. Beijing: Beijing University of Posts and Telecommunications,2019.
宋晓敏.基于改进贝叶斯算法的中文信息分类研究[D].北京:北京邮电大学,2019.

[14] KOHAVI R. Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid[C]. Proceedings of KDD, 1996:202-207.

[15] LE M M. Reserach and application of data mining classification algorithm[D]. Chengdu: University of Electronic Science and Technology of China,2017.
乐明明.数据挖掘分类算法的研究和应用[D].成都:电子科技大学,2017.

[16] LI X. A comparative study on five decision tree algorithms[D]. Dalian: Dalian University of Technology,2011.
李旭.五种决策树算法的比较研究[D].大连:大连理工大学,2011.

[17] HAN J W, KAMBER M. Data mining:concepts and techniques[M]. Translated by FAN M,MENG X F. Beijing: Beijing: China Machine Press,2007.
HAN J W, KAMBER M. 数据挖掘:概念与技术[M].范明,孟小峰,译.北京:机械工业出版社,2007.

[18] MIU L F. Research and application of improved C4.5 algorithm in college students' emotional quality analysis[D]. Shanghai: Shanghai Normal University,2018.
缪连芬.改进的C4.5算法在大学生情感素质分析中的研究与应用[D].上海:上海师范大学,2018.

[19] LI J S. Research on Bayesian classifier based on volume test[D]. Beijing: Beijing Jiaotong University,2008.
李锦善.基于Volume Test的贝叶斯分类器研究[D].北京:北京交通大学,2008.

[20] ZHOU Z H. Machine learning[M]. Beijing: Tsinghua University Press,2016.
周志华.机器学习[M].北京:清华大学出版社,2016.

(责任编辑:孙　娟)