影评分类的朴素贝叶斯文本分类算法改进

张浩强[1]，任思行[1]，曾繁慧[2]，刘海涛[2]
（1. 辽宁工业大学 机械工程与自动化学院，辽宁 锦州，121001;
2. 辽宁工程技术大学 理学院，辽宁 阜新，123000）

摘要：针对影评分类的情感分析这种文本分类问题中朴素贝叶斯分类算法的局限性，根据句法依存关系在文本中抽取情感特征，将其向量化；利用主成分分析降低自变量的相关性；通过样本数据训练得到优化的朴素贝叶斯分类器；优化模型对余下的样本数据预测及与朴素贝叶斯算法进行对比.康奈尔大学网站共享的影视评价数据集研究结果表明：改进的朴素贝叶斯分类算法大大提高了准确率.

关键词：朴素贝叶斯；主成分分析；文本分析；影评分类
中图分类号：O 212.8

Improved Naive Bayes Text Classification Algorithm for Film Classification

ZHANG Haoqiang[1], REN Sihang[1], ZENG Fanhui[2], LIU Haitao[2]
(1. College of Mechanical Engineering and Automation,Liaoning University of Technology,Jinzhou 121001,China;
2. College of Science,Liaoning Technical University,Fuxin 123000,China)

Abstract: The sentiment analysis of film classification belongs to the problem of text classification. In order to solve the limitation of the Naive Bayes classification algorithm, according to the syntactic dependency relation, and the emotional features are extracted from the text and then to quantify it. Using principal component analysis to reduce the correlation of independent variables, and the optimized Bayes classifier is obtained by training the sample data. The optimal model is used to predict the remaining data and compare with the Naive Bayes algorithm. The research results of video evaluation data set shared by Cornell University show that: Improved Naive Bayesian classification algorithm greatly improves the accuracy.

Key words: Naive Bayes；Principal component analysis; The text analysis；Film classification

0 引言

网络技术的发展产生了大量的文本数据，这些文本数据数量巨大导致简单的处理办法无法直接获得有效的信息.如何快速有效的提取出大数据中具有价值的知识，是人们一直关注的热点问题.朴素贝叶斯分类算法是以文本分词为目的，以贝叶斯公式为原理的一种数据挖掘方法.文献[1]扩展了朴素贝叶斯分类器的应用，提高了朴素贝叶斯分类器的分类精度.文献[2]提出一种基于粗糙集的特征加权朴素贝叶斯算法(FWNB).文献[3]提出了一种基于相似度的实例加权的朴素贝叶斯分类算法和一种基于 C4.5 和 NB 的组合分类算法.文献[4]提出了基于 M-估计的贝叶斯分类算法(FISC-M)和加权集成的贝叶斯分类算法(WFISC).

本文对影视评价的文本数据进行数据挖掘.线上影视业的在线评论量超过百万，一部电影、一家公司，如何更加人性化的改善，影迷评论是重中之重.大数据中，利用朴素贝叶斯分类算法来计算影视的好评率是非常快捷且准确的，但是由于朴素贝叶斯分类算法要求变量之间相互独立，而实际生活中的数据很难满足这个条件，因此本文利用主成分分析方法降低

基金项目：辽宁省教育厅科学技术研究基金项目（L2014133）
作者简介：张浩强(1995-),男，学士，主要研究方向：数据挖掘
通信联系人：刘海涛（1982-），男，讲师，主要研究方向：数据挖掘. E-mail: 597873883@qq.com

变量之间的相关性，得到优化的朴素贝叶斯文本分类改进算法，以此来提高算法的准确性.基于 Python 平台，分别用改进的优化模型与朴素贝叶斯分类算法对康奈尔大学网站共享的影评数据集进行分类预测，以检验算法的准确率.

1 影评分类的朴素贝叶斯文本分类算法

1.1 朴素贝叶斯文本分类算法

假设对于某个数据集，随机变量 C 表示样本为 C 类的概率，F_1 表示测试样本某特征出现的概率，套用基本贝叶斯公式

$$P(C \mid F_1)=\frac{P(CF_1)}{P(F_1)}=\frac{P(C)\cdot P(F_1 \mid C)}{P(F_1)} \quad . \qquad (1\text{-}1)$$

式（1-1）表示对于某个样本，特征 F_1 出现时，该样本被分为 C 类的条件概率[5-6].测试样本分类：

假设有 n 个测试样本，其特征 F_1 出现了（F_1=1），那么就计算 $P(C=0 \mid F_1=1)$和 $P(C=1 \mid F_1=1)$的概率值.前者大，则该样本被认为是 0 类；后者大，则分为 1 类.

对于多个特征而言，贝叶斯公式可以扩展为

$$\begin{aligned} P(C \mid F_1F_2\cdots F_n) &= \frac{P(C)\cdot P(F_1F_2\cdots F_n \mid C)}{P(F_1F_2\cdots F_n)} \\ &= \frac{P(C)\cdot P(F_1 \mid C)\cdot P(F_2\cdots F_n \mid CF_1)}{P(F_1F_2\cdots F_n)} \\ &= \cdots \\ &= \frac{P(C)\cdot P(F_1 \mid C)\cdot P(F_2 \mid CF_1)\cdots P(F_n \mid CF_1\cdots F_{n-1})}{P(F_1F_2\cdots F_n)} \end{aligned} \qquad (1\text{-}2)$$

这就是著名的贝叶斯定理，朴素贝叶斯分类器是一个以贝叶斯定理为基础，广泛应用于情感分类领域的优美分类器.

对于扩展后的贝叶斯公式（1-2），分子中存在一大串似然值.当特征很多的时候，这些似然值的计算不方便.为了简化计算，朴素贝叶斯算法假设：朴素的认为各个特征相互独立[6].如此一来，式（1-2）的分子就简化为

$$P(C) * P(F1|C) * P(F2|C) \ldots P(Fn|C).$$

这个假设认为各个特征之间是独立的，看上去确实是个很不科学的假设.因为很多情况下，各个特征之间是紧密联系的.然而在朴素贝叶斯的大量应用实践表明其工作的相当好.

其次，由于朴素贝叶斯的工作原理是计算$P(C=0|F1\ldots Fn)$和$P(C=1|F1\ldots Fn)$，并取最大值的那个作为其分类.而二者的分母是一模一样的.因此，又可以省略分母计算，从而进一步简化计算过程.

另外，贝叶斯公式推导能够成立有个重要前期，就是各个证据（evidence）不能为 0.也即对于任意特征$Fx, P(Fx)$不能为 0.而显示某些特征未出现在测试集中的情况是可以发生的.因此实现上通常要做一些小的处理，例如把所有计数进行+1，即加法平滑(additive smoothing，又叫拉普拉斯平滑(Laplace smothing).而如果通过增加一个大于 0 的可调参数 alpha 进行平滑，就叫 Lidstone 平滑.

例如，在所有 6 个分为 C=1 的影评样本中，某个特征 F_1=1 不存在，则 $P(F_1=1|C=1)$ = 0/6，

$P(F_1=0|C=1) = 6/6$.

经过加法平滑后，$P(F_1=1|C=1) = (0+1)/(6+2)=1/8$，$P(F_1=0|C=1) = (6+1)/(6+2)=7/8$.

注意分母的+2，这种特殊处理使得 2 个互斥事件的概率和恒为 1.

当特征很多的时候，大量小数值的小数乘法会有溢出风险.因此，通常的实现都是将其转换为

$$\log[P(C)\times P(F_1|C)\times P(F_2|C)\times\cdots\times P(F_n|C)] = \log[P(C)]+\log[P(F_1|C)] +\cdots+\log[P(F_n|C)]$$

将乘法转换为加法，就彻底避免了乘法溢出风险.

1.2 实例分析

本文使用的数据集共有 2 个标签，一个为“net”，一个为“pos”，每个目录下面有 5 个文本文件.目录如表 1.

表 1 **文本样本分布**

Tab.1 Text sample distribution

negetive	positive
sb movie.	worth it!
a shit movie.	worth my money.
waste of time.	a nb movie.
shit.	I love this movie!
waste my money.	nb!

（1）文本特征

从这些英文中抽取情感态度而进行分类，最直观的做法就是抽取单词.通常认为，很多关键词能够反映说话者的态度.例如上面这个简单的数据集，很容易发现，凡是说了“shit”的，就一定属于 neg 类.

当然，上面数据集是为了方便描述而简单设计的.现实中一个词经常会有模棱两可的态度.但是仍然有理由相信，某个单词在 neg 类中出现的越多，那么表示 neg 态度的概率越大.

同样注意到有些单词对情感分类是毫无意义的.比如上述数据中的“of”，“I”之类的单词.这类词有个名字，叫“Stop_Word“（停用词）.这类词是可以完全忽略掉不做统计的.显然忽略掉这些词，词频记录的存储空间能够得到优化，而且构建速度也更快.

把每个单词的词频作为重要的特征也存在一个问题.比如上述数据中的”movie“，在 10 个样本中出现了 5 次，但是出现正反两边次数差不多，没有什么区分度.而”worth“出现了 2 次，但却只出现在 pos 类中，显然更具有强烈的刚晴色彩，即区分度很高.

因此，需要引入 TF-IDF（Term Frequency-Inverse Document Frequency，词频和逆向文件频率）对每个单词做进一步考量[7].

TF（词频）的计算很简单，就是针对一个文件 t，某个单词 Nt 出现在该文档的频率[8].例如文档“I love this movie”，单词“love”的 TF 为 1/4.如果去掉停用词“I"和”it“，则为 1/2.

IDF（逆向文件频率）的意义是，对于某个单词 t，凡是出现了该单词的文档数 Dt，占了全部测试文档 D 的比例，再求自然对数[9].

比如单词“movie“一共出现了 5 次，而文档总数为 12，因此 IDF 为 ln(5/12).

很显然，IDF 是为了凸显那种出现的少，但是占有强烈感情色彩的词语.比如“movie”这样的词的 IDF=ln(12/5)=0.88，远小于“love”的 IDF=ln(12/1)=2.48.

TF-IDF 就是把二者简单的乘在一起即可.这样，求出每个文档中，每个单词的 TF-IDF，就是提取得到的文本特征值.在训练和预测的时候为了保证其量化，将 net 设置成 0，pos 设置成 1.

经过 TF-IDF 对上述 10 个文本文件处理之后，得到如下矩阵，见表 2，其中每一行代表一个文本本件，每一列代表某个单词在文本文件中的权重.

表 2 文本样本向量化

Tab.2 Text sample to quantify

	love	money	movie	nb	sb	shit	time	waste	worth	n/p
sb movie	0	0	0.544	0	0.839	0	0	0	0	0
a shit movie	0	0	0.609	0	0	0.79	0	0	0	0
waste of time	0	0	0	0	0	0	0.764	0.645	0	0
shit	0	0	0	0	0	1	0	0	0	0
waste my money	0	0.707	0	0	0	0	0	0.707	0	0
worth it	0	0	0	0	0	0	0	0	1	1
worth my money	0	0.707	0	0	0	0	0	0	0.707	1
a nb movie	0	0	0.544	0.84	0	0	0	0	0	1
i love this movie	0.84	0	0.544	0	0	0	0	0	0	1
nb	0	0	0	1	0	0	0	0	0	1

（2）朴素贝叶斯分类结果

利用朴素贝叶斯分类算法对表 2 的矩阵进行训练和预测，总样本的 80%来训练朴素贝叶斯分类模型，然后利用训练好的模型对剩余 20%进行测试，最后再与真实值进行对比，得到结果见表 3.

表 3 小样本预测值与真实值对比

Tab.3 Small sample comparison between predictive value and true value

准确值	预测值
1	0
1	1

根据表 3 计算可以得出，一般的朴素贝叶斯分类算法准确率为：0.50.

2 影评分类的朴素贝叶斯文本分类算法改进

在人们提出朴素贝叶斯分类算法的时候，为了便于计算，提出了一个非常大胆的假设：因变量之间相互独立.然后朴素贝叶斯分类算法成为了文本分类的大热门，因为词语之间的相关性真的很小.

但是，像下面这种例子：

“老师，您的学识很渊博，我很敬佩您.”

可以看出，在“渊博”和“敬佩”之间还是存在着很大的相关性，这就和朴素贝叶斯分类原理的基本假设起了冲突.因此设想找出一种用来降低变量之间相关性的算法，用其对朴素贝叶斯算法进行优化，使分类预测的准确率提高.基于这个设想，本文根据主成分分析的特点，利用主成分分析降低变量之间的相关性，再用朴素贝叶斯分类算法进行分类预测.

2.1 主成分分析

实际问题的样本数据往往包含多个指标变量，这些指标之间常常存在着一定程度的、有时甚至是相当高的相关性，使观测数据中的信息在一定程度上有所重迭. 这种多重相关性的危害很多，实际问题的系统分析必须消除变量的多重相关性，尽量减少重叠信息的不良作用. 主成分分析是克服变量多重相关性的一种基本方法，是消解多维随机变量各个分量间线性相关性以及变量系统降维的基本方法. 主成分分析通过降维技术把多个指标约化为少数几个综合指标，这些综合指标能够反映原始指标的绝大部分信息，它们通常表示为原始指标的某种线性组合.

主成分分析主要步骤如下：

（1）对原始数据进行标准化；

（2）计算相关系数矩阵；

（3）计算特征值与特征向量；

（4）计算主成分贡献率及累计贡献率；

（5）确定主成分个数；

（6）计算主成分载荷；

（7）各主成分得分.

2.2 改进的贝叶斯文本分类算法

文本特征抽取过程省略，利用 1.2 实例分析中的数据，将文本样本向量化，得到矩阵见表 2.

对表 2 向量化之后的样本作主成分分析，首先将样本数据进行标准化，然后计算相关系数矩阵，得到特征值与特征向量.计算过程省略.

特征值中选取 7 个主成分，总占比为 99%.每个成分的占比见表 4.

表 4 各个成分的占比

Tab.4 Proportion of each component

	占比
V_1	0.228
V_2	0.185
V_3	0.152
V_4	0.132
V_5	0.078
V_6	0.075
V_7	0.025
V_8	0.01
V_9	0

主成分分析后的主成分矩阵见表 5.

表 5 主成分矩阵

Tab.5 Principal component matrix

	Pc_1	Pc_2	Pc_3	Pc_4	Pc_5	Pc_6	Pc_7

sb movie	0.29	-0.05	0.02	0.61	-0.59	-0.02	0.1
a shit movie	0.48	-0.57	0.07	-0.13	0	-0.07	-0.26
waste of time	-0.3	0	-0.8	-0.02	0	0.49	-0.06
shit	0.28	-0.68	0.04	-0.51	0	0	0.21
waste my money	-0.6	0.01	-0.5	-0.03	0	-0.47	0
worth it	-0.6	0.01	0.62	0	0	0.37	-0.02
worth my money	-0.7	0.01	0.38	-0.01	0	-0.24	-0.04
a nb movie	0.49	0.58	0.06	-0.1	0	-0.06	-0.23
i love this movie	0.29	-0.05	0.02	0.61	0.59	-0.02	0.1
nb	0.31	0.73	0.04	-0.41	0	0	0.19

接下来利用朴素贝叶斯分类算法对主成分分析之后样本做分类预测，同样的 80%做训练样本，来训练朴素贝叶斯分类模型，然后用训练好的模型对剩余的 20%样本进行分类预测，得出混淆矩阵见表 6.由表 6 可以看出，准确率为 100%.

表 6 小样本预测混淆矩阵

Tab.6 Small sample prediction confusion matrix

		预测值	
		0	**1**
真实值	**0**	1	0
	1	0	1

3 算法对比

为对比算法，用大样本进行测试，所用数据为康纳尔大学网站共享影评数据集.将数据集分为两类：positive 和 negetive 类，每一类包括 700 个文件，从中每一类抽取 20 个文件，每个文件大约 600 个单词.同样，利用其中的 80%作为训练样本，来训练朴素贝叶斯分类模型，然后用剩余的 20%来进行预测，以得出模型的准确率.

（1）朴素贝叶斯分类算法

首先使用 TF-IDF 算法进行分词，把文本信息转化成可以数字矩阵的形式，得到 5 371 个情感词，然后使用未优化之前的朴素贝叶斯分类对其进行分类预测，依然用 80%的样本作为训练样本，来训练朴素贝叶斯分类器，然后用训练好的分类器对剩余的 20%的样本进行预测，最后与真实值进行对比，经计算可以得出，其准确率为：0.625.

表 7 大样本预测值与真实值对比

Tab.7 Large sample comparison between predictive value and true value

预测值	真实值
1	1
0	0
0	0
1	1
0	0
0	1
0	1
0	1

（2）改进的朴素贝叶斯分类算法

接下来对 TF-IDF 处理生成的数字矩阵进行主成分分析，选取 34 个主成分，主成分占比为 97%，然后再用朴素贝叶斯分类算法对其进行分类预测.同样的 80%的样本作为训练样本，来训练朴素贝叶斯分类器，然后用训练好的分类器对剩余的 20%的样本进行预测，最

后与真实值进行对比.

混淆矩阵见表 8.

表 8 大样本混淆矩阵

Tab.8 Large sample prediction confusion matrix

		预测值	
		0	**1**
真实值	**0**	3	1
	1	0	4

由表 8 混淆矩阵可以看出，经主成分分析优化之后，改进的朴素贝叶斯分类算法的准确率为：0.875，比未优化前的朴素贝叶斯分类算法大有提高.

4 结论

在数据时代发展飞速的今天，数据量的爆炸性增加，使得人类处理起来非常困乏，所以利用机器学习来处理大规模数据.本文利用主成分分析对朴素贝叶斯分类算法进行优化，得到改进的贝叶斯分类模型，经由大量的影视评论样本进行验证，改进模型的分类准确率大大提高.本文思想希望可以对机器学习的发展提供些许助力.

[参考文献] (References)

[1] [1] 李静梅,孙丽华,张巧荣,等.一种文本处理中的朴素贝叶斯分类器[J].哈尔滨工程大学学报,2015,10:456-470.
[2] 王国才.朴素贝叶斯分类器的研究与应用[D].重庆交通大学,2010.
[3] 蒋良孝.朴素贝叶斯分类器及其改进算法研究[D].中国地质大学,2009.
[4] 段晶.朴素贝叶斯分类及其应用研究[D].大连海事大学,2011.
[5] YAN Zhiyong,XU Congfu,PAN Yunhe.Improving naive Bayes classifier by dividing its decision regions[J].Journal of Zhejiang University-Science C:Computers & Electronics,2011 (8):119-139.
[6] 范焱,郑诚,王清毅,等.用 Naive Bayes 方法协调分类 Web 网页[J].软件学报,2001(9):33-60.
[7] 张建娥.基于 TFIDF 和词语关联度的中文关键词提取方法[J].情报科学,2012(10): 33-35.
[8] 施聪莺,徐朝军,杨晓江.TFIDF 算法研究综述[J].计算机应用,2009(S1):47-53.
[9] 张玉芳,彭时名,吕佳.基于文本分类 TFIDF 方法的改进与应用[J].计算机工程,2006(19):56-65.