Advances in Applied Mathematics 应用数学进展, 2024, 13(6), 2901-2911
Published Online June 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.136278

Hans汉斯

基于距离相关系数的局部实例加权朴素贝叶斯文本分类算法

骆洁琴[1]，彭　萍[2]，胡桂开[1]

[1]东华理工大学理学院，江西　南昌
[2]东华理工大学经济与管理学院，江西　南昌

收稿日期：2024年5月28日；录用日期：2024年6月22日；发布日期：2024年6月28日

摘　要

朴素贝叶斯算法具有简单高效的特点，被广泛应用于文本分类。方法要求属性之间满足条件独立性假设，然而该假设在现实中很难满足。同时，随着大数据时代到来，文本数据呈现非线性结构的特点，经典朴素贝叶斯算法拟合效果不高。为解决以上问题，本文提出了一种基于距离相关系数的局部实例加权朴素贝叶斯分类算法。首先，计算属性和类别的距离相关系数，并将其作为属性权重嵌入到文档距离测度中，构建一种新的距离度量方法；其次，测算训练样本和测试样本的距离，进行实例选择和实例加权，构建局部实例加权贝叶斯文本分类器；最后，利用WEKA平台上的15个文本数据集对算法性能进行实验比较。结果表明新提出的算法在分类精度上均优于三种经典的朴素贝叶斯文本分类器。

关键词

文本分类，朴素贝叶斯，实例选择，实例加权，距离相关系数

A Locally Instance Weighting Naive Bayes Text Classification Algorithm Based on Distance Correlation Coefficient

Jieqin Luo[1], Ping Peng[2], Guikai Hu[1]

[1]School of Sciences, East China University of Technology, Nanchang Jiangxi
[2]School of Economics and Management, East China University of Technology, Nanchang Jiangxi

Received: May 28th, 2024; accepted: Jun. 22nd, 2024; published: Jun. 28th, 2024

文章引用: 骆洁琴, 彭萍, 胡桂开. 基于距离相关系数的局部实例加权朴素贝叶斯文本分类算法[J]. 应用数学进展, 2024, 13(6): 2901-2911. DOI: 10.12677/aam.2024.136278

Abstract

Naive Bayes algorithm has the characteristics of simplicity and efficiency, and is widely used in text classification. The method requires the assumption of conditional independence between attributes, which is difficult to satisfy in reality. Meanwhile, with the advent of the big data era, text data exhibits non-linear structures, and the fitting effect of classical naive Bayesian algorithms is limited. To address these issues, a locally instance-weighted Naive Bayes classification algorithm based on distance correlation coefficient is proposed. Firstly, it calculates the distance correlation coefficient between attributes and classes, and embeds it as attribute weights into the document distance measure to construct a new distance measurement method. Secondly, it measures the distances between training samples and test samples, conducts instance selection and instance weighting, and constructs a locally instance-weighted Bayesian text classifier. Finally, the algorithm's performance is experimentally compared with 15 text datasets from the WEKA platform. The results indicate that the proposed algorithm outperforms three classical Naive Bayes text classifiers in terms of classification accuracy.

Keywords

Text Classification, Naive Bayes, Instance Selection, Instance Weighting, Distance Correlation Coefficient

Copyright © 2024 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

随着科学技术的迅速发展，信息数据呈爆炸式增长，如何从各类海量文本数据中挖掘信息成为人们面临的巨大挑战，文本分类作为其中的关键技术，受到广泛研究。目前，常用的文本分类算法包括：多项式朴素贝叶斯[1]、决策树[2]、支持向量机[3]、神经网络[4]等。其中，多项式朴素贝叶斯是文本分类的主流算法之一，具有计算简单、高效的特点。多项式朴素贝叶斯要求给定文档分类情形下属性之间相互独立，但在实际问题中，这个假设很难成立，从而影响分类精度。为削弱独立性假设对算法性能的影响，考虑到同一个类别下的每个文档对分类的贡献程度是不一样的，学者们从实例加权方向提出了许多改进算法。

实例加权是通过某个评估方法对每个实例赋予不同的权重，体现实例间的贡献差异。如何构建实例权重，是实例加权改进算法的关键。目前，部分学者从不同角度提出了改进，主要包含两个方向：全局实例加权和局部实例加权。全局实例加权是对所有实例赋权后再进行分类。如，Jiang [5]等将每个实例的权重设置为 1，再根据估计的条件概率与真实的条件概率之间的差值迭代更新权重，提出了判别加权的朴素贝叶斯。Xu [6]等利用实例的属性值频率向量与属性值个数向量的内积定义实例的权重，提出了基于属性值频率加权的朴素贝叶斯。局部实例加权是实例选择和实例加权相结合的算法，首先根据各个文档与测试文档的距离选择 k 个训练文档，然后利用距离反比函数对 k 个文档赋权。相对于全局实例加权，局部实例加权算法效率较高，得到学者们高度重视。如，Frank [7]等利用 k 近邻算法寻找与测试文档最近的 k 个训练文档，根据与测试文档的距离对每个训练文档进行赋权，提出了局部加权的朴素贝叶

斯。Jiang [8]等基于互信息构造了两个不同文本的距离计算公式，通过这个距离公式选出与测试文档距离最近的部分样本，在选出的样本上进行局部实例加权。

以上实例加权算法比没有加权的多项式朴素贝叶斯算法，分类精度有所提高，但随着文本数据呈现出来的特征，如高维稀疏，大体量，非线性结构等，实例加权多项式朴素贝叶斯算法还存在很大的改进空间。因此，本文提出一种基于距离相关系数的局部实例加权朴素贝叶斯文本分类器(Locally Instance Weighted Naive Bayes Text Classifiers, LIWNB)，旨在提高文本数据的分类效果。论文主要创新点如下：

1) 利用距离相关系数构建属性权重，解决属性和类别维度不同无法直接计算相关系数的问题，测度属性间非线性依赖关系。

2) 将距离相关系数权重嵌入到距离度量公式中，认为属性对分类的贡献是不相等的，强化属性间的差异。

3) 基于局部实例加权进行文本分类，克服文本数据高维稀疏、体量大的特点，优化算法效率。

论文结构安排如下：第 2 节介绍常用的三种朴素贝叶斯文本分类器，并给出距离相关系数的定义；第 3 节详细介绍基于距离相关系数改进的局部实例加权朴素贝叶斯分类器算法；第 4 节基于 WEKA 平台上数据集对算法的性能进行实验比较；第 5 节为论文结论。

2. 预备知识

2.1. 朴素贝叶斯文本分类器

文本数据进行分类前，需要对文本进行表示，常见的表示方法包括基于向量空间的模型[9]，潜在语义模型[10]和概率生成模型[11]等。朴素贝叶斯文本分类模型是在向量空间模型上构建的，包括基于二项分布的伯努利模型和基于多项分布的多项式朴素贝叶斯、补集朴素贝叶斯和两者的结合模型。伯努利模型将文档由空间中的二进制向量表示，不考虑单词出现的频数，该模型在小数据集上效果良好，但遇到大型数据集时，分类精度会受到影响。为解决伯努利模型的缺点，多项式贝叶斯模型(Multinomial Naive Bayes, MNB)被提出，该模型统计了单词在文档中的频数。即一篇测试文档可以表示为一个向量 $d=(w_1,w_2,\cdots,w_m)$， w_i 表示文档中的第 i 个单词。在属性条件独立的假设前提下，MNB 使用下式对文档 d 进行分类：

$$c_{MNB}(d)=\arg\max_{c\in C}P(c)\prod_{i=1}^{m}P(w_i\mid c)^{f_i} \tag{1}$$

其中 f_i 是单词 w_i 在文档 d 中出现的频数。对上述公式取对数不会影响分类精度，因此上式可进一步简化为：

$$c_{MNB}(d)=\arg\max_{c\in C}\left[\log P(c)+\sum_{i=1}^{m}f_i\log P(w_i\mid c)\right] \tag{2}$$

其中，先验概率 $P(c)$ 和条件概率 $P(w_i\mid c)$ 分别由以下公式表示

$$P(c)=\frac{\sum_{j=1}^{n}\delta(c(j),c)+1}{n+l} \tag{3}$$

$$P(w_i\mid c)=\frac{\sum_{j=1}^{n}f_{ji}\delta(c(j),c)+1}{\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ji}\delta(c(j),c)+m} \tag{4}$$

其中 n 为训练文档的数量，l 为类别的种类数，m 为属性个数，$c(j)$ 表示第 j 篇文档对应的类别，f_{ji} 表示第 j 篇文档中第 i 个单词出现的频数。

当其中一个类别比其它类别拥有多很多的训练文档时，训练文档数量较少的类别的权重会变小，导致MNB的分类精度下降。因此，提出了补集贝叶斯模型(Complement Naive Bayes, CNB)，CNB对文档 d 的分类公式如下：

$$c_{CNB}(d)=\arg\max_{c\in C}\left[-\log P(\bar{c})-\sum_{i=1}^{m}f_i\log P(w_i\mid\bar{c})\right] \tag{5}$$

其中 $\bar{c}$ 是类别 c 的补集(除了类别 c 以外的所有类)，先验概率和条件概率的计算公式如下：

$$P(\bar{c})=\frac{\sum_{j=1}^{n}\delta(c(j),\bar{c})+1}{n+l} \tag{6}$$

$$P(w_i\mid\bar{c})=\frac{\sum_{j=1}^{n}f_{ji}\delta(c(j),\bar{c})+1}{\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ji}\delta(c(j),\bar{c})+m} \tag{7}$$

(One-versus-all-but-one, OVA)是MNB与CNB的直接结合，它使用公式(8)对文档 d 进行分类：

$$c_{OVA}(d)=\arg\max_{c\in C}\left[\left(\log P(c)-\log P(\bar{c})\right)+\sum_{i=1}^{m}f_i\left(\log P(w_i\mid c)-\log P(w_i\mid\bar{c})\right)\right] \tag{8}$$

其中概率 $P(c)$、$P(w_i\mid c)$、$P(\bar{c})$、$P(w_i\mid\bar{c})$ 分别用公式(3)、(4)、(6)、(7)计算。

2.2. 距离相关系数

相关系数可以用于判别两个随机向量之间的线性关系，相关系数越小，表示线性关系越弱，但当相关系数为0时并不能说明两变量是独立的。因此，Szekely [12]等学者提出了距离相关系数，它通过两个随机变量的联合特征函数与各自的边际特征函数乘积之间的差来描述相关程度，由于特征函数能唯一确定随机变量的概率分布。距离相关系数不仅可以度量两个变量之间的线性关系，也能度量变量间的非线性关系，且当距离相关系数为0时，可以说明两个变量是独立的。距离相关系数被广泛应用于许多领域。如，Miao [13]根据距离相关系数对不同规模的变量进行聚类，发现所需时间更短，效率更高。学者孙[14]利用距离相关系数融合GPR模型，评估各遥测参数和响应变量之间的相关关系，成功降低了卫星异常检测的虚警率。Bhattacharjee [15]利用距离相关系数观测血压和血清胆固醇之间的相关性，并将其应用于临床数据分析。本文拟将其应用到贝叶斯分类算法中，接下来给出样本距离相关系数的定义。

定义1 [16]：假设 $(X,Y)=\{(X_k,Y_k):k=1,2,\cdots,n\}$ 是随机向量 (X,Y) 的观测样本，则样本距离协方差 $V_n(X,Y)$ 定义为

$$V_n^2(X,Y)=\frac{1}{n^2}\sum_{k,l=1}^{n}A_{kl}B_{kl} \tag{9}$$

其中 $A_{kl}=a_{kl}-\overline{a_{k\cdot}}-\overline{a_{\cdot l}}+\overline{a_{\cdot\cdot}}$，$k,l=1,2,\cdots,n$。$a_{kl}=\left|X_k-X_l\right|_p$，$\overline{a_{k\cdot}}=\frac{1}{n}\sum_{l=1}^{n}a_{kl}$，$\overline{a_{\cdot l}}=\frac{1}{n}\sum_{k=1}^{n}a_{kl}$，$\overline{a_{\cdot\cdot}}=\frac{1}{n^2}\sum_{k,l=1}^{n}a_{kl}$。同样地，$B_{kl}=b_{kl}-\overline{b_{k\cdot}}-\overline{b_{\cdot l}}+\overline{b_{\cdot\cdot}}$，$b_{kl}=\left|Y_k-Y_l\right|_p$，$k,l=1,2,\cdots,n$。

当 $X=Y$ 时，距离方差的定义为：

$$V_n^2(X)=V_n^2(X,Y)=\frac{1}{n^2}\sum_{k,l=1}^{n}A_{kl}^2 \tag{10}$$

定义 2 [17]：假设$(X,Y)=\{(X_k,Y_k):k=1,2,\cdots,n\}$是随机向量$(X,Y)$的观测样本，在随机向量 X 与 Y 一阶矩有限的情况下，则样本的距离相关系数 $R_n(X,Y)$ 定义为

$$R_n^2(X,Y)=\begin{cases}\dfrac{V_n^2(X,Y)}{\sqrt{V_n^2(X)V_n^2(Y)}}, & V_n^2(X)V_n^2(Y)>0\\ 0, & V_n^2(X)V_n^2(Y)=0\end{cases} \tag{11}$$

3. 基于距离相关系数改进的局部实例加权朴素贝叶斯分类器

局部加权方法是一种实例选择方法，它通过筛选部分训练样本构建分类器，算法的核心问题是构建合适的距离度量对训练样本进行选择。本节拟结合局部加权方法改进朴素贝叶斯文本分类器，具体步骤如下：

首先，计算文本数据中单词和类别的距离相关系数，得到各个单词的权重。假设每个文本可以由 m 个单词构成，在第 j 篇文本中第 i 个属性 w_i 出现的频数为 $f_{ji}(j=1,2,\cdots,n;i=1,2,\cdots,m)$，则第 i 个属性 w_i 在所有文本中的取值可构成列向量 $F_i=(f_{1i},f_{2i},\cdots,f_{ni})$，类别记为$C=(c_1,c_2,\cdots,c_m)$，共 m 类。由定义 1 和 2 可知，随机变量样本 $(F_i,C)=\{(f_{ji},c(j)):j=1,2,\cdots,n\}$ 的距离协方差 $V_n(F_i,C)$ 和距离相关系数 $D_n(F_i,C)$ 的计算[18]如下：

$$V_n^2(F_i,C)=\frac{1}{n^2}\sum_{k,l=1}^{n}A_{kl}B_{kl} \tag{12}$$

$$D_n^2(F_i,C)=\begin{cases}\dfrac{V_n^2(F_i,C)}{\sqrt{V_n^2(F_i)V_n^2(C)}}, & V_n^2(F_i)V_n^2(C)>0\\ 0, & V_n^2(F_i)V_n^2(C)=0\end{cases} \tag{13}$$

其中，$A_{kl}=a_{kl}-\overline{a_{k\cdot}}-\overline{a_{\cdot l}}+\overline{a_{\cdot\cdot}}$，$a_{kl}=|f_{ki}-f_{li}|$，$\overline{a_{k\cdot}}=\frac{1}{n}\sum_{l=1}^{n}a_{kl}$，$\overline{a_{\cdot l}}=\frac{1}{n}\sum_{k=1}^{n}a_{kl}$，$\overline{a_{\cdot\cdot}}=\frac{1}{n^2}\sum_{k,l=1}^{n}a_{kl}$，$k,l=1,2,\cdots,n$。$B_{kl}=b_{kl}-\overline{b_{k\cdot}}-\overline{b_{\cdot l}}+\overline{b_{\cdot\cdot}}$，$b_{kl}=1-\delta(c(k),c(l))$，$k,l=1,2,\cdots,n$。由于文本数据中的类别是名义变量，在实验中，当属性属于同一类别时，b_{kl} 设置为 0，属于不同类别时，b_{kl} 设置为 1。

其次，将距离相关系数 $D_n(F_i,C)$ 嵌入到两个文档的距离公式中。将两篇文档 x 与 y 之间的距离定义为：

$$dis(x,y)=\sqrt{\sum_{i=1}^{m}\left[D_n(F_i,C)\frac{f_i(x)-f_i(y)}{f_i(x)+f_i(y)+1}\right]^2} \tag{14}$$

$f_i(x)$ 是文档 x 中第 i 个单词出现的次数，$f_i(y)$ 是文档 y 中第 i 个单词出现的次数。距离(14)是一种改进加权距离。距离中的 $D_n(F_i,C)$ 是各个属性的权重，体现了各个属性的差异，认为每个属性的重要性是不相等的，这也符号实际数据情形。$\frac{f_i(x)-f_i(y)}{f_i(x)+f_i(y)+1}$ 使距离取值在[0, 1]之间，消除量纲的影响，同时对分母进行平滑化处理，克服了数据稀疏问题。

然后，由公式(14)计算测试文档 d 与所有训练文档的距离，选出与测试文档距离最近的 k 个样本，并对这 k 个样本赋予权重，与测试文档较近的训练实例赋予较大的权重，越远的训练实例赋予越小的权重。测试文档第 j 个邻居 d_j 的权重 W_j 计算公式为：

$$W_j = \frac{1}{1+dis\left(d,d_j\right)^2} \tag{15}$$

最后，在选出的 k 个训练样本中构建实例加权的朴素贝叶斯文本分类器。在后续实验中，本文使用改进后的MNB、CNB、OVA分类器进行实验，分别用LIWMNB、LIWCNB、LIWOVA表示。在使用改进后的模型时，决策函数分别类似于(2)、(5)、(8)，唯一区别是先验概率与条件概率的计算，在本文的LIWNBTC算法中，公式(3)、(4)、(6)、(7)被公式(16)、(17)、(18)、(19)替代。具体算法流程见算法1。

$$P\left(c\right) = \frac{\sum_{j=1}^{k} W_j \delta\left(c\left(j\right),c\right)+1}{\sum_{j=1}^{k} W_j + l} \tag{16}$$

$$P\left(w_i \mid c\right) = \frac{\sum_{j=1}^{k} W_j f_{ji} \delta\left(c\left(j\right),c\right)+1}{\sum_{i=1}^{m}\sum_{j=1}^{k} W_j f_{ji} \delta\left(c\left(j\right),c\right)+m} \tag{17}$$

$$P\left(\overline{c}\right) = \frac{\sum_{j=1}^{k} W_j \delta\left(c\left(j\right),\overline{c}\right)+1}{\sum_{j=1}^{k} W_j + l} \tag{18}$$

$$P\left(w_i \mid \overline{c}\right) = \frac{\sum_{j=1}^{k} W_j f_{ji} \delta\left(c\left(j\right),\overline{c}\right)+1}{\sum_{i=1}^{m}\sum_{j=1}^{k} W_j f_{ji} \delta\left(c\left(j\right),\overline{c}\right)+m} \tag{19}$$

算法 1 LIWNBTC 算法

输入： 训练文档集 D，一个测试文档 d,邻域的大小 k

输出： d 的预测类标签

1) 用公式(14)计算测试文档 d 和每个训练文档 d_j 之间的距离 $dis\left(d,d_j\right)$
2) 找出最近的 k 个邻居 $d_1,d_2,\cdots,d_k$
3) 初始化局部实例加权训练文档集 $\{d_1,d_2,\cdots,d_k\}$
4) 对于每个邻居 d_j, $j=1,2,\cdots,k$，使用公式(15)对其设置权重 W_j
5) 在局部训练文档集 $\{d_1,d_2,\cdots,d_k\}$ 上构建朴素贝叶斯文本分类器
6) 使用构建的朴素贝叶斯文本分类器预测测试文档 d 的类标签
7) 返回测试文档 d 的类标签

局部训练文档集的大小是使用 k 近邻算法确定的，通过实验验证发现，LIWNBTC 对 k 的选择不敏感，这与 Frank [7]得出的结论一致。对不同的 k 值，分类效果差别不大，k 一般不小于 30，因此在后续实验中，将 k 值设为 30。由于不同的测试文档的邻域是不同的，因此需要对每一个测试文档 d 都计算一次概率。

4. 实验与分析

4.1. 实验环境与数据

为了评估算法 LIWNBTC 的分类性能，我们开展了三组实验，分别比较 LIWMNB、LIWCNB、

LIWOVA 与 MNB、CNB 和 OVA 在分类精度方面的差异。实验均是在 Anaconda3 环境下，采用 python3.10 编程语言，电脑系统是 windows11，内存 16 G 的电脑上完成。

实验数据选自国际平台 WEKA 提供的 15 个带标签的标准文本分类数据集，详细的数据介绍如表 1 所示。

Table 1. Text classification dataset used in the experiment
表 1. 实验使用的文本分类数据集

数据集	属性个数	实例个数	类别数
fbis	2000	2463	17
oh0	1003	3182	10
oh10	1050	3238	10
oh15	913	3100	10
oh5	918	3012	10
re0	1657	3758	25
re1	1504	2886	13
tr11	414	6429	9
tr12	313	5804	8
tr21	336	7902	6
tr23	204	5832	6
tr31	927	10,128	7
tr41	878	7454	10
tr45	690	8261	10
wap	1560	8460	20

4.2. 实验结果分析

实验中，每个算法在每个数据集上的分类精度都是通过十次十折交叉验证获得的，不同的算法在相同的训练集和测试集上进行运行并评估。表 2~4 分别给出了各个分类器在各数据集上的分类精度，表中符号“◆”表示对应列的算法获得显著性的优势。表格中的数据分别表示与原始朴素贝叶斯文本分类器相比，改进算法的精度。在表格最底部，总结了每个算法的平均分类精度和输赢比(w/t/l)，w/t/l 表示改进的朴素贝叶斯文本算法与原始的朴素贝叶斯文本算法相比，改进的算法在 w 个数据集上获胜，在 t 个数据集上持平，l 个数据集上失败。

为了进一步了解结果，进行了配对双尾 t 检验，显著性水平设置为 0.05 来比较每对算法。详细的比较结果见表 5 和表 6。表 5 中的每个数字表示这列算法与这行算法相比，获得显著性胜利的数据集的数量。表 6 中，第一列是算法名称，第二列的数字是对应行的算法与其它所有算法相比总共获得的胜利的数据集的数量和总共失败的数据集的数量之差，该列可用于生成排名。第三列和第四列分别表示对应行的算法胜利和失败的数据集的总数。

从实验结果可以看出，改进的朴素贝叶斯文本算法明显优于原始的朴素贝叶斯文本算法。总结如下：

1) LIWMNB 以 9 胜 5 输的成绩优于 MNB，LIWCNB 以 8 胜 5 输的成绩优于 CNB，LIWOVA 以 8 胜 6 输的成绩优于 OVA。

2) CNB 以 8 胜 2 输的成绩优于 MNB，OVA 以 9 胜 0 输的成绩远远优于 MNB。

3) 从分类准确率排序的表 6 中可以看出，CNB 算法获得胜利的总数据集数与失败的总数据集数的差值是最大的，在分类精度排名上，CNB 是优于 MNB 和 OVA 的，这一结论与 Rennie [19]得出的结论一致，他们也发现 CNB 的性能优于 MNB 和 OVA。

Table 2. Experimental results of MNB and LIWMNB: classification accuracy and standard deviation
表 2. MNB 与 LIWMNB 的实验结果：分类精度和标准差

数据集	MNB	LIWMNB
fbis	77.06 ± 0.025	84.23 ± 0.02◆
oh0	89.78 ± 0.029	83.99 ± 0.034
oh10	80.49 ± 0.036	77.44 ± 0.046
oh15	83.50 ± 0.039	79.75 ± 0.041
oh5	86.48 ± 0.035	87.22 ± 0.036
re0	79.75 ± 0.03	83.31 ± 0.034◆
re1	83.25 ± 0.032	85.96 ± 0.026◆
tr11	84.58 ± 0.06	89.18 ± 0.047◆
tr12	81.03 ± 0.067	83.29 ± 0.06◆
tr21	61.51 ± 0.074	86.24 ± 0.063◆
tr23	70.52 ± 0.102	78.05 ± 0.089◆
tr31	94.46 ± 0.022	96.91 ± 0.016◆
tr41	94.67 ± 0.024	93.20 ± 0.023
tr45	83.09 ± 0.048	85.49 ± 0.042◆
wap	80.99 ± 0.032	68.37 ± 0.035
Average	80.99	84.18
w/t/l	-	9/1/5

Table 3. Experimental results of CNB and LIWCNB: classification accuracy and standard deviation
表 3. CNB 与 LIWCNB 的实验结果：分类精度和标准差

数据集	CNB	LIWCNB
fbis	76.79 ± 0.027	84.36 ± 0.023◆
oh0	92.17 ± 0.029	86.22 ± 0.036
oh10	81.69 ± 0.04	76.67 ± 0.045
oh15	84.28 ± 0.039	79.30 ± 0.043

续表

oh5	90.83 ± 0.029	86.69 ± 0.03
re0	82.61 ± 0.029	84.00 ± 0.025◆
re1	84.84 ± 0.028	86.20 ± 0.024◆
tr11	82.12 ± 0.061	89.53 ± 0.04◆
tr12	85.53 ± 0.063	84.06 ± 0.066
tr21	85.73 ± 0.088	87.84 ± 0.064◆
tr23	69.09 ± 0.11	78.22 ± 0.084◆
tr31	94.70 ± 0.024	97.40 ± 0.017◆
tr41	94.18 ± 0.023	94.31 ± 0.025
tr45	87.10 ± 0.034	89.43 ± 0.032◆
wap	77.54 ± 0.03	71.17 ± 0.033
Average	84.61	85.03
w/t/l	-	8/2/5

Table 4. Experimental results of OVA and LIWOVA: classification accuracy and standard deviation
表 4. OVA 与 LIWOVA 的实验结果：分类精度和标准差

数据集	OVA	LIWOVA
fbis	80.86 ± 0.025	84.59 ± 0.023◆
oh0	91.20 ± 0.026	84.55 ± 0.037
oh10	81.78 ± 0.035	77.88 ± 0.033
oh15	84.10 ± 0.041	79.97 ± 0.039
oh5	89.13 ± 0.034	87.54 ± 0.039
re0	81.12 ± 0.028	83.57 ± 0.027◆
re1	84.68 ± 0.027	86.14 ± 0.026◆
tr11	85.73 ± 0.054	89.80 ± 0.05◆
tr12	83.57 ± 0.059	84.06 ± 0.064
tr21	71.38 ± 0.087	86.79 ± 0.053◆
tr23	71.16 ± 0.106	79.04 ± 0.093◆
tr31	94.78 ± 0.023	97.21 ± 0.018◆
tr41	95.03 ± 0.023	93.85 ± 0.026
tr45	85.83 ± 0.044	87.04 ± 0.038◆
wap	80.36 ± 0.03	69.21 ± 0.04
Average	84.05	84.75
w/t/l	-	8/1/6

Table 5. Classification accuracy test (p = 0.05)
表 5. 分类准确率检验(p = 0.05)

	MNB	CNB	OVA	LIWMNB	LIWCNB	LIWOVA
MNB	-	8	9	9	9	9
CNB	2	-	4	5	8	6
OVA	0	6	-	7	8	8
LIWMNB	5	8	6	-	5	1
LIWCNB	4	5	6	0	-	1
LIWOVA	5	5	6	0	3	-

Table 6. Classification accuracy ranking (p = 0.05)
表 6. 分类准确率排序(p = 0.05)

数据集	赢－输	赢	输
LIWCNB	17	33	16
CNB	7	32	25
LIWOVA	6	25	19
LIWMNB	2	31	29
OVA	−4	21	25
MNB	−28	16	44

5. 结语

本文首先回顾了经典的三种朴素贝叶斯文本分类器，然后将基于距离相关系数改进的局部实例加权方法与朴素贝叶斯文本分类器结合，提出一种新的局部实例加权朴素贝叶斯文本分类算法，实验表明LIWMNB、LIWCNB、LIWOVA 在分类精度上分别优于 MNB、CNB、OVA，达到了削弱条件独立性假设，解决非线性结构的文本分类问题的目的。

改进算法的精确度有明显提高，但由于距离相关系数的计算导致了更高的时间复杂度，原因在于距离相关系数是基于全部训练样本得到的，并且针对所有属性构建距离度量。为此，后期我们拟尝试将实例选择和属性选择相结合的思想对算法进行改进，提高算法的效率。

基金项目

国家自然科学基金(11661003)；江西省自然科学基金(20192BAB201006)。

参考文献

[1] McCallum, A. and Nigam, K. (1998) A Comparison of Event Models for Naive Bayes Text Classification. In: *Proceedings of the* 15*th AAAI Workshop on Learning for Text Categorization* (*AAAI*'98). AAAI Press/The MIT Press, Madison, Wisconsin, 41-48.

[2] Hall, M. (2007) A Decision Tree-Based Attribute Weighting Filter for Naive Bayes. *Knowledge-Based Systems*, **20**, 120-126. https://doi.org/10.1016/j.knosys.2006.11.008

[3] Joachims, T. (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In:

Nédellec, C. and Rouveirol, C., Eds., *Machine Learning*: *ECML*-98, Springer, 137-142. https://doi.org/10.1007/BFb0026683

[4] Sebastiani, F. (2002) Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, **34**, 1-47. https://doi.org/10.1145/505282.505283

[5] Jiang, L., Wang, D. and Cai, Z. (2012) Discriminatively Weighted Naive Bayes and Its Application in Text Classification. *International Journal on Artificial Intelligence Tools*, **21**, Article ID: 1250007. https://doi.org/10.1142/s0218213011004770

[6] Xu, W., Jiang, L. and Yu, L. (2018) An Attribute Value Frequency-Based Instance Weighting Filter for Naive Bayes. *Journal of Experimental & Theoretical Artificial Intelligence*, **31**, 225-236. https://doi.org/10.1080/0952813x.2018.1544284

[7] Frank, E., Hall, M. and Pfahringer, B. (2003) Locally Weighted Naive Bayes. arXiv: 1212.2487.

[8] Jiang, L., Cai, Z., Zhang, H. and Wang, D. (2013) Naive Bayes Text Classifiers: A Locally Weighted Learning Approach. *Journal of Experimental & Theoretical Artificial Intelligence*, **25**, 273-286. https://doi.org/10.1080/0952813x.2012.721010

[9] Salton, G., Wong, A. and Yang, C.S. (1975) A Vector Space Model for Automatic Indexing. *Communications of the ACM*, **18**, 613-620. https://doi.org/10.1145/361219.361220

[10] Ababneh, A.H., Lu, J. and Xu, Q. (2019) An Efficient Framework of Utilizing the Latent Semantic Analysis in Text Extraction. *International Journal of Speech Technology*, **22**, 785-815. https://doi.org/10.1007/s10772-019-09623-8

[11] Su, J., Zeng, J., Xiong, D., Liu, Y., Wang, M. and Xie, J. (2018) A Hierarchy-To-Sequence Attentional Neural Machine Translation Model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **26**, 623-632. https://doi.org/10.1109/taslp.2018.2789721

[12] Székely, G.J., Rizzo, M.L. and Bakirov, N.K. (2007) Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, **35**, 2769-2794. https://doi.org/10.1214/009053607000000505

[13] Miao, C. (2021) Clustering of Different Dimensional Variables Based on Distance Correlation Coefficient. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-020-02817-y

[14] 孙宇豪, 李国通, 张鸽. 距离相关系数融合 GPR 模型的卫星异常检测方法[J]. 北京航空大学学报, 2021, 47(4): 844-852.

[15] Bhattacharjee, A. (2014) Distance Correlation Coefficient: An Application with Bayesian Approach in Clinical Data Analysis. *Journal of Modern Applied Statistical Methods*, **13**, 354-366. https://doi.org/10.22237/jmasm/1398918120

[16] Sheng, W. and Yin, X. (2016) Sufficient Dimension Reduction via Distance Covariance. *Journal of Computational and Graphical Statistics*, **25**, 91-104. https://doi.org/10.1080/10618600.2015.1026601

[17] Li, R., Zhong, W. and Zhu, L. (2012) Feature Screening via Distance Correlation Learning. *Journal of the American Statistical Association*, **107**, 1129-1139. https://doi.org/10.1080/01621459.2012.695654

[18] Ruan, S., Chen, B., Song, K. and Li, H. (2021) Weighted Naïve Bayes Text Classification Algorithm Based on Improved Distance Correlation Coefficient. *Neural Computing and Applications*, **34**, 2729-2738. https://doi.org/10.1007/s00521-021-05989-6

[19] Rennie, J., Shih, L., Teevan, J. and Karger, D. (20003) Tackling the Poor Assumptions of Naive Bayes Text Classifiers. *Proceedings of the Twentieth International Conference on Machine Learning* (*ICML*-2003), Washington DC, 21-24 August 2003, 616-623.