Advances in Applied Mathematics 应用数学进展, 2023, 12(10), 4300-4309
Published Online October 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.1210423

Hans汉斯

类依赖实例加权朴素贝叶斯算法研究

曾嘉琪，彭　萍*，杨　柳，胡桂开

东华理工大学理学院，江西　南昌

收稿日期：2023年9月17日；录用日期：2023年10月10日；发布日期：2023年10月17日

摘　要

为削弱朴素贝叶斯中属性条件独立性假设的影响，人们提出了许多改进朴素贝叶斯的方法，实例加权是一个重要的改进方向，但现有实例权重构造是将训练样本作为一个整体进行处理，没有考虑类内实例的分布情况。因此，本文提出两种类依赖实例加权朴素贝叶斯算法：基于相关性的类依赖实例加权朴素贝叶斯(CCSIWNB)和类依赖属性值频率实例加权朴素贝叶斯(CSAVFWNB)。关于CCSIWNB，实例权重是在计算类内每个实例与该类众数实例相似度后，消除该实例与其它类众数实例的平均相似度基础上得到的。关于CSAVFWNB，实例的权重是由类内属性值频率向量和该类属性值个数向量的内积得到的。最后，采用标准UCI数据集将CCSIWNB、CSAVFWNB与朴素贝叶斯算法和其它实例加权朴素贝叶斯算法进行仿真实验，结果表明本文提出的算法在准确率上优于其它算法。

关键词

朴素贝叶斯，实例加权，类依赖，分类器

Study on Class-Specific Instance Weighted Naive Bayes

Jiaqi Zeng, Ping Peng*, Liu Yang, Guikai Hu

School of Sciences, East China University of Technology, Nanchang Jiangxi

Received: Sep. 17th, 2023; accepted: Oct. 10th, 2023; published: Oct. 17th, 2023

Abstract

In order to weaken the influence of attribute conditional independence hypothesis in naive Bayes, many improved naive Bayes methods have been proposed, and instance weighting is an important improvement direction. However, the existing instance weight construction considers the training sample as a whole, without considering the distribution of instances in the class. Therefore, two kinds of class-specific weighted naive Bayes algorithms are proposed in this paper: correlation-

*通讯作者。

文章引用：曾嘉琪，彭萍，杨柳，胡桂开. 类依赖实例加权朴素贝叶斯算法研究[J]. 应用数学进展，2023，12(10): 4300-4309. DOI: 10.12677/aam.2023.1210423

based class-specific instance weighted Naive Bayes (CCSIWNB) and class-specific attribute value frequency instance weighted Naive Bayes (CSAVFWNB). About CCSIWNB, the weight is obtained on the basis of calculating the similarity between each instance of certain class and the mode instance of the same class, and eliminating the average similarity between the instance and the mode instances of the other class. For CSAVFWNB, the weight of each instance is the inner product of the attribute value frequency vector and the attribute value number vector in the same class. Finally, CCSIWNB and CSAVFWNB are simulated with naive Bayes algorithm and other case weighted naive Bayes algorithm using standard UCI data set. The results show that the proposed algorithm is superior to other algorithms in accuracy.

Keywords

Naive Bayes, Case Weighting, Class-Specific, Classifier

1. 引言

朴素贝叶斯是数据挖掘领域中常用的一种分类算法，模型不但简单，而且在解决各种问题中效果良好，且不易受噪声影响，非常稳健。在分类学习过程，分类器给每个新实例分配一个类标签，朴素贝叶斯利用贝叶斯定理计算新实例最可能的标签。假设测试实例 x 对应的属性值为 $a_1,a_2,\cdots,a_m$，m 是属性的个数，a_j 是 x 的第 j 个属性值，则实例 x 被划分为具有最大后验概率的类别，即 $c(x)=\arg\max_{c\in C} P(c)P(a_1,a_2,\cdots,a_m\mid c)$，其中 C 是类别 c 所有可能值的集合。在朴素贝叶斯学习中，假设给定类下所有属性都是独立的，即 $P(a_1,a_2,\cdots,a_m\mid c)=\prod_{j=1}^{m}P(a_j\mid c)$，因此最大后验概率分类为

$$c(x)=\arg\max_{c\in C} P(c)\prod_{j=1}^{m}P(a_j\mid c), \tag{1}$$

然而，属性条件独立性假设在现实生活中往往难以达到，学者们提出了不同方面的改进，大致分为两个方向：一是基于不切实际的属性条件独立性假设；二是对不可靠的概率估计进行改进。

方向一包括结构扩展和属性加权(选择)两种方法。结构扩展主要研究属性间的关联关系，主要有网状结构的朴素贝叶斯，树扩展朴素贝叶斯，平均单依赖估测器模型，隐朴素贝叶斯等结构扩展，具体可参考文献[1]-[6]。属性加权是对每个属性赋予不同的权重，属性选择是从原始属性集中删除不具有预测能力或预测能力微弱的属性，即对入选属性赋予权重 1，删除属性赋予权重 0，所以属性选择是一种特殊的属性加权，具体可参阅文献[7] [8] [9] [10]。众所周知，为了避免概率为零，公式(1)中的先验概率 $P(c)$ 和条件概率 $P(a_j\mid c)$ 利用拉普拉斯平滑方法计算：

$$P(c)=\frac{\sum_{i=1}^{n}\delta(c_i,c)+1}{n+k}, \tag{2}$$

$$P(a_j\mid c)=\frac{\sum_{i=1}^{n}\delta(a_{ij},a_j)\delta(c_i,c)+1}{\sum_{i=1}^{n}\delta(c_i,c)+n_j}, \tag{3}$$

其中 n 为训练实例的个数，k 为类的个数，n_j 为第 j 个属性值的个数，c_i 为第 i 个训练实例的类标号，a_{ij} 为第 i 个训练实例的第 j 个属性值，$\delta(a,b)$ 是示性函数，当 $a=b$ 时，函数值为 1，当 $a\neq b$ 时，函数值为 0。

方向二是对不可靠的概率估计进行改进，也就是对公式(2)和(3)进行改进，具体包含两种方法：概率微调和实例加权(选择)。概率微调是先利用 NB 分类，然后根据每个训练分类结果对先验概率(3)进行微调，再利用微调后的概率进行 NB 分类，从而提高分类的准确率，具体可参阅文献[11] [12] [13] [14] [15]。实例加权是根据训练实例的分布或者重要程度给每个实例分配不同的权重，是先计算出所有训练实例的权重，再利用加权后的数据来构建朴素贝叶斯分类器，即先给每个实例赋予不同的权重计算概率(2)和(3)，再进行 NB 分类。相对概率微调法，它不仅对条件概率计算优化，同时对先验概率进行改进，而概率微调法主要对条件概率进行改进。此时公式(2)和(3)改进为：

$$P(c)=\frac{\sum_{i=1}^{n} w_i\delta(c_i,c)+1}{\sum_{i=1}^{n} w_i+k}, \tag{4}$$

$$P(a_j\mid c)=\frac{\sum_{i=1}^{n} w_i\delta(a_{ij},a_j)\delta(c_i,c)+1}{\sum_{i=1}^{n} w_i\delta(c_i,c)+n_j}, \tag{5}$$

其中 w_i 是第 i 个训练实例的权重。

实例加权是在实例的维度上对数据进行处理，实例选择是一种特殊的实例加权，当实例的权重为 1 时，实例被保留，当权重为 0 时，实例被删除。实例加权(选择)是当下研究的一个热门方向，如何确定权重 w_i 是问题研究的关键。考虑到不同实例与测试实例的远近程度对分类的影响也不同，往往通过计算测试实例与训练实例之间的距离来确定实例权重，如 Xie 等[16]提出了一种基于选择性邻域的惰性学习朴素贝叶斯算法，其基本思想是根据测试实例与训练实例之间属性值不同的个数将训练实例划分为不同半径的领域，在不同领域上构建多个朴素贝叶斯分类器，利用分类精度最高的分类器对测试实例分类；Frank 等[17]提出了局部加权朴素贝叶斯算法，该算法将局部加权线性回归思想运用到朴素贝叶斯的实例加权中，实例的权重根据训练实例与测试实例之间的欧氏距离计算得到，离测试实例近的训练实例分配更多的权重。另外，考虑到实例中各属性取值的分布情况对分类的影响不一样，通过分析各实例属性分布情况确定实例权重，如 Jiang [18]等通过计算每个实例和实例众数之间的相似度确定实例权重，通过衡量每个实例与中心实例的接近程度提出了改进实例加权朴素贝叶斯算法(Instance Weighted Naive Bayes, IWNB)；Xu [19]等通过计算属性值频率向量与属性值个数向量的内积作为实例的权重，提出了基于属性值频率加权朴素贝叶斯(Attribute Value Frequency-Based Instance Weighted Naive Bayes, AVFWNB)。杨等[20]考虑到众数实例非唯一和每个属性对分类的影响不一样，利用算术平均和加权平均，将属性权重嵌入到实例权重的构造中，对 IWNB 算法进行改进，提出了嵌入属性加权的实例加权朴素贝叶斯算法(Embedded-attribute Weighted Instance-weighted Naive Bayes, EAWIWNB)。

对于上述实例加权算法，不管是基于距离还是属性值分布构造权重，它们都是把所有实例作为一个整体进行考虑，没有考虑到类别对权重构造的影响。因此，本文在基于属性值分布情况的基础上将类别考虑进去，在不同类别下考虑属性值分布情况构造实例权重提出了两种新的类依赖实例加权算法，分别是基于相关性的类依赖实例加权朴素贝叶斯算法和类依赖属性值频率加权朴素贝叶斯算法，并利用 UCI 数据集进行实证分析。

下文结构安排如下：第 2 节介绍了基于相关性的类依赖实例加权朴素贝叶斯；第 3 节介绍了类依赖属性值频率加权朴素贝叶斯；第 4 节详细描述了实验设置和实验结果；第 5 节对全文进行总结。

2. 基于相关性的类依赖实例加权朴素贝叶斯

实例加权中关键的步骤是如何给每个实例分配权重，IWNB 算法是通过计算每个实例和实例众数之间的相似度来确定实例的权重，它是根据属性取值分布情况来确定权重的，相对 NB 而言，在分类精度上有所提高。然而，对方考虑的属性取值分布情况是对整个训练样本集中考虑，没有与类别同时考虑。众所周知，训练样本都有给定的类标签，每个实例对分类的重要性和它的类标签密切相关的，所以在考虑实例中各个属性取值的分布情况时，要结合实例的类别进行考虑，每个实例与它所在类的众数实例，即类中心的相似度要大，与其它类中心的相似度要小。换句话说，对于每个实例，我们要肯定其对自身所在类的贡献，同时要消除其对其它类的影响。为此，我们参照 IWNB 算法的思想，结合类别分析属性取值的分布情况，同时消除类间影响(在其他类别中出现造成的冗余信息)，提出了一种基于相关性的类依赖实例加权朴素贝叶斯算法(Correlation-Based Class-Specific Instance Weighted Naive Bayes, CCSIWNB)，该算法在构造实例权重时不仅考虑到了类别来优化不可靠的概率估计，还通过消除冗余信息来弱化属性间条件独立性的影响。CCSIWNB 算法首先根据训练样本的分类，将不同样本进行归类；其次参照 Jiang [18]给出的属性值众数、实例众数定义得到每类的实例众数；然后通过判断训练实例所属类别，并利用重叠度量方法计算出实例与其所属类别实例众数的相似度，并减去与其他类众数实例相似度的平均值，最后利用 Sigmoid 函数归一化得到该实例的权重，具体计算公式如下：

$$R_i = s\left(x_i, y_j\right) - \frac{1}{k-1}\sum_{t=1\wedge t\neq j}^{k} s\left(x_i, y_t\right) \tag{6}$$

$$w_i = \frac{1}{1+\mathrm{e}^{-R_i}}, \tag{7}$$

其中 x_i 为训练集的第 i 个实例，y_j 表示第 j 类实例众数，k 为类别个数，$s\left(x_i, y_j\right) = \sum_{t=1}^{m}\delta\left(a_t\left(x_i\right), a_t\left(y_j\right)\right)$ 表示第 i 个实例与第 j 类实例众数的相似度，$\sum_{t=1\wedge t\neq j}^{k} s\left(x_i, y_t\right)$ 表示第 i 个实例与其他类实例众数的相似度的和，通过求平均得到该实例与其它类间的平均冗余度，公式(7)为归一化处理，w_i 为第 i 个实例的最终权重。CCSIWNB 算法流程见下表 1：

Table 1. CCSIWNB algorithm flow
表 1. CCSIWNB 算法流程

算法 1：CCSIWNB
输入：训练实例集 D，测试实例 x_{test}
输出：测试实例 x_{test} 的类标签
Step1：将训练实例集 D 按类标签进行划分
Step2：计算训练实例集 D 的所有类实例众数 y 以及与训练实例的相似度
Step3：利用公式(6)和(7)给每个训练实例赋予权重
Step4：利用加权后的 D 建立朴素贝叶斯模型
Step5：通过模型预测测试实例 x_{test} 的类标签
Step6：输出 x_{test} 的类标签

3. 类依赖属性值频率加权朴素贝叶斯

IWNB 算法利用重叠度量方法计算训练实例与众数实例的相似度构造实例权重，获得的实例权重都为整数，数值上缺乏权重的含义，只考虑了属性值是否同时出现，没有利用属性值的频率信息，为此 Xu 等[19]学者提出了一种基于属性值频率的实例加权分类器，该方法考虑了两个方面：1)属性值的频率包含一些重要的信息，这些信息也可以用来定义训练实例的权重。2) 每个训练实例的权值与其属性值频率向量和整个训练数据集的属性值个数向量呈正相关。与 IWNB 相比，AVFWNB 充分考虑了每个属性值的频率信息，同时涉及更大的解空间，即实例权重从整数范围扩大到了连续的正有理数，但该方法没有考虑到类别对实例权重构造的影响。

为解决以上问题，本文提出了一种基于属性值频率的类依赖实例加权朴素贝叶斯(Class-specific Attribute Value Frequency Instance Weighted Naive Bayes, CSAVFWNB)。该算法首先将训练数据按类别进行划分，接着对每个类计算出类依赖属性值频率 $f_{c,ij}$ 以及类依赖属性值个数 $n_{c,j}$，最后考虑按类划分之后的类依赖属性值频率向量与类依赖属性值个数向量呈正相关，将两者做内积后得到的数值作为实例权重，具体计算公式如下：

$$f_{c,ij}=\frac{\sum_{k=1}^{n}\delta\left(a_{kj},a_{ij}\right)\delta\left(c_k,c_i\right)}{n_{c_i}} \tag{8}$$

其中 $f_{c,ij}$ 为 a_{ij} (第 i 个训练实例的第 j 个属性值)的类依赖属性值频率，c_i,c_k 分别为第 i 个实例与第 k 个实例的类别标签，n_{c_i} 为第 i 个实例所在类的实例总个数，a_{kj} 为第 k 个训练实例的第 j 个属性值。

另外，设 $n_{c,j}$ 为第 c 类中第 j 个属性的属性值个数，则类依赖属性值数向量可表示为 $\left(n_{c,1},n_{c,2},\cdots,n_{c,m}\right)$。然后，根据实例权重确定的方法，第 i 个训练实例权重定义为其类依赖属性值频率向量与类依赖属性值数量向量的内积(标量积)。具体计算公式为：

$$w_i=\left(f_{c,i1},f_{c,i2},\cdots,f_{c,im}\right)\cdot\left(n_{c,1},n_{c,2},\cdots,n_{c,m}\right)=\sum_{j=1}^{m}\left(f_{c,ij}*n_{c,j}\right) \tag{9}$$

w_i 为最终的类依赖实例权重，CSAVFWNB 算法流程见表 2：

Table 2. CSAVFWNB algorithm flow
表 2. CSAVFWNB 算法流程

算法 2：CSAVFWNB
输入：训练实例集 D，测试实例 x_{test}
输出：测试实例 x_{test} 的类标签
Step1：将训练实例集 D 按类标签进行划分
Step2：利用公式(8)计算每个属性值 a_{ij} 的类依赖属性值频率 $f_{c,ij}$
Step3：利用公式(9)计算出每个实例的类依赖实例权重 w_i
Step4：利用公式(4)和(5)分别计算 P_c 和 $P\left(a_j\mid c\right)$
Step5：利用模型预测测试实例 x_{test} 的类标签
Step6：输出 x_{test} 的类标签

4. 实验结果与分析

4.1. 实验环境与数据处理

实验环境：本文所有实验都是在 win10 系统，128G(SSD) + 1T(HDD)硬盘，Intel(R) Core(TM) i5-6300HQ CPU，内存 16G 的 PC 机上通过 R 4.2.0 版本完成。

本文在 UCI 数据集中选取了 10 个具有分类标签的数据集进行实验，见表 3。这些数据集来源于现实生活的多个领域，包含各种类型的数据特征。我们首先对数据集进行预处理，读取实例个数、属性个数以及分类个数，判断是否存在缺失值，对于包含缺失值的案例，实验中采用删除该案例的方式进行处理。

Table 3. 10 experimental data sets
表 3. 10 个实验数据集

数据集名称	实例个数	特征个数	分类个数	是否存在缺失值	是否存在名义变量
Iris	150	4	3	N	N
Tae	151	5	3	N	N
Balance.scale	625	4	3	N	N
Breast_Cancer_Coimbra	116	9	2	N	N
Breast.cancer	286	9	2	Y	Y
Breast.cancer.wisconsin	699	9	2	Y	N
Somerville.happiness.survey	143	6	2	N	N
Crx	690	15	2	Y	Y
Seeds_dataset	210	7	3	N	N
Reprocessed.hungarian	294	13	5	N	N
Car	1728	6	4	N	Y

4.2. 结果与分析

实验结果是通过将数据集随机分成 90%的训练集及 10%的测试集，最后进行 10 折交叉验证得到结果进行平均而获得的。表 4 给出了四种算法分类精度的详细比较结果，第一列为实验所用测试数据集，后面几列为四种算法在这些数据集下的平均分类精确度以及方差，其中标记有“√”和“*o*”的符号表示当进行 p = 0.05 显著性水平的配对双尾 t 检验时，以第一种算法为参照，其他算法是否与其存在显著差异，“√”和“*o*”分别表示该算法在某个数据集上明显优于或输于第一种算法，平均值和输赢(W/T/L)值汇总在表格底部。

Table 4. Accuracy comparison of CCSIWNB with NB, IWNB, EAWIWNB
表 4. CCSIWNB 与 NB、IWNB、EAWIWNB 的精确度比较

数据集名称	四种算法精确度平均值			
	CCSIWNB	NB	IWNB	EAWIWNB
Iris	94.99 ± 4.71	93.99 ± 4.71	94.33 ± 4.84	93.66 ± 4.84
Tae	78.67 ± 8.05	78.67 ± 8.05	84.00 ± 7.12○	78.67 ± 7.12

Continued

Balance.scale	93.39 ± 2.49	92.58 ± 3.44	91.61 ± 3.00	90.48 ± 4.50√
Breast_Cancer_Coimbra	96.36 ± 5.30	96.36 ± 5.30	95.45 ± 5.38	95.00 ± 5.93
Breast.cancer	81.79 ± 6.28	74.29 ± 4.01√	76.61 ± 5.81√	75.67 ± 8.24√
Breast.cancer.wisconsin	96.67 ± 2.87	91.30 ± 3.97√	92.00 ± 4.92√	96.96 ± 3.30
Somerville.happiness.survey	61.51 ± 16.95	62.23 ± 17.64	61.52 ± 14.87	58.38 ± 17.34√
Crx	89.28 ± 6.83	80.18 ± 4.17√	80.27 ± 5.86√	82.27 ± 6.98√
Seeds_dataset	67.22 ± 16.47	68.18 ± 14.87	70.63 ± 13.38	67.81 ± 12.10
Reprocessed.hungarian	66.03 ± 11.66	62.96 ± 11.09√	64.36 ± 11.59	65.10 ± 6.39
Average	82.59	80.07	81.08	80.4
W/T/L	-	0/6/4	1/6/3	0/6/4

Wilcoxon 符号秩检验是一种常用的非参数统计检验，它适用检验成对观测数据之差是否来自均值为 0 的总体(产生数据的总体是否具有相同的均值)，对每个数据集的算法对的性能差异进行排序，同时忽略符号，并比较正负差异的秩。表 5 总结了 CCSIWNB 与其他三种算法的比较结果，其中 V 值表示基于两种算法在不同数据集下每次交叉验证得到的精确度差值且符号为正号的秩和。p 值表示检验统计量大于或等于实际观察样本数据计算得到的检验统计量值的概率，符号“√”表示在该数据集下 CSCIWNB 和 CSAVFWNB 算法优于所比较的算法。

Table 5. Comparison results of Wilcoxon signed rank test between CSCIWNB and other algorithms
表 5. CSCIWNB 与其他算法的 Wilcoxon 符号秩检验比较结果

数据集名称	CCSIWNB 与 NB		CCSIWNB 与 IWNB		CCSIWNB 与 EAWIWNB	
	V	p-value	V	p-value	V	p-value
Iris	25	0.3543	10	0.5862	101	0.2534
Tae	10	0.08897	2	0.002552√	132.5	0.1361
Balance.scale	7	0.5839	49	0.03109√	87	0.0323√
Breast_Cancer_Coimbra	12	0.7921	20	0.8028	81	0.4773
Breast.cancer	66	0.00370√	82.5	0.01051√	141.5	0.0341√
Breast.cancer.wisconsin	89.5	0.00218√	65	0.004366√	72	0.04853√
Somerville.happiness.survey	0	1	17	0.944	72.5	0.836
Crx	131	0.00122√	160	0.001265√	151	0.08936
Seeds_dataset	7	0.5271	14	0.3411	81	0.8497
Reprocessed.hungarian	6	0.1814	44	0.1016	116	0.4093

从表 4 与表 5 的比较结果可以看出，CCSIWNB 算法明显优于 NB，而且优于 IWNB 和 EAWIWNB 等实例加权算法，具体结果总结如下：

1) CCSIWNB 在 10 个数据集上的平均分类准确率为 82.59%，显著高于 NB (80.07%)以及其他两种实例加权算法 IWNB (81.08%)，EAWIWNB (80.4%)。

2) 基于配对双尾 t 检验，CCSIWNB 优于 NB (4 胜 6 平 0 负)、IWNB (3 胜 6 平 1 负)，EAWIWNB (4 胜 6 平 0 负)。

3) 根据 Wilcoxon 符号秩检验结果，我们可以得出结论，当水平显著性为 $\alpha = 0.05$ 时，CCSIWNB 显著优于 NB、IWNB 和 EAWIWNB。

Table 6. Accuracy comparison of CSAVFWNB with NB, IWNB and AVFWNB
表 6. CSAVFWNB 与 NB、IWNB、AVFWNB 的精确度比较

数据集名称	四种算法精确度平均值			
	CSAVFWNB	NB	IWNB	AVFWNB
Iris	94.2 ± 10.19	94.82 ± 7.49	94.82 ± 9.34	94.82 ± 7.49
Tae	84.29 ± 13.06	78.5 ± 19.2√	84.2 ± 15.12	73.87 ± 13.82√
Balance.scale	93.06 ± 2.84	92.58 ± 3.44	92.25 ± 3.00	92.5 ± 4.16
Breast_Cancer_Coimbra	95.96 ± 5.23	93.64 ± 5.30	95.45 ± 5.38	95.00 ± 5.93
Breast.cancer	76.64 ± 6.88	77.62 ± 9.56	76.53 ± 9.33	76.47 ± 9.16
Breast.cancer.wisconsin	97.14 ± 2.39	94.30 ± 4.57√	94.00 ± 4.72√	93.96 ± 2.34√
Somerville.happiness.survey	60.98 ± 17.7	58.57 ± 17.84	56.6 ± 19.43√	59.38 ± 17.34
Crx	84.37 ± 6.36	81.18 ± 5.67√	80.27 ± 6.46√	77.07 ± 7.45√
Seeds_dataset	65.5 ± 15.9	66.45 ± 14.66	70.27 ± 11.53○	71.22 ± 11.68
Reprocessed.hungarian	65.01 ± 7.97	65.94 ± 11.59	67.03 ± 10.5	65.54 ± 10.01
Average	81.72	80.36	81.14	79.98
W/T/L	-	0/7/3	1/6/3	1/6/3

Table 7. Comparison results of Wilcoxon signed rank test between CSAVFWNB and other algorithms
表 7. CSAVFWNB 与其他算法的 Wilcoxon 符号秩检验比较结果

数据集名称	CSAVFWNB 与 NB		CSAVFWNB 与 IWNB		CSAVFWNB 与 AVFWNB	
	V	p-value	V	p-value	V	p-value
Iris	2	0.7728	0	1	2	0.7728
Tae	51	0.01188√	13.5	1	72	0.01064√
Balance.scale	11	1	35	0.04752√	73.5	0.7954
Breast_Cancer_Coimbra	6	0.1736	1	1	18	0.1198
Breast.cancer	55.5	0.8195	45.5	0.637	53	0.6243
Breast.cancer.wisconsin	0	1	0	0.3458	3.5	0.3222
Somerville.happiness.survey	24.5	0.03991√	24	0.01058√	14	0.5271
Crx	12	0.02353√	84	0.04935√	167.5	0.003837√
Seeds_dataset	27	0.6212	30	0.088	14	0.04982√
Reprocessed.hungarian	29.5	0.7894	24	0.1403	50.5	0.7527

从表 6 与表 7 的比较结果可以看出，CSAVFWNB 算法明显优于 NB，而且优于 IWNB 和 AVFWNB 等算法，实验结果总结如下：

CSAVFWNB 在 10 个数据集上的平均分类准确率为 81.72%显著高于 NB (80.36%)，IWNB (81.14%)，AVFWNB (79.92%)。

基于配对双尾 t 检验，CSAVFWNB 优于 NB (3 胜 7 平 0 负)、IWNB (3 胜 6 平 1 负)，AVFWNB (3 胜 6 平 1 负)。

根据 Wilcoxon 符号秩检验结果，我们可以得出结论，当水平显著性为 $\alpha = 0.05$ 时，我们提出的 CSAVFWNB 显著优于 NB、IWNB 和 AVFWNB。

5. 结论

为了弱化不切实际的属性条件独立性假设以及改进不可靠的概率估计，本文提出了两种类依赖实例加权方法，CCSIWNB 算法不仅考虑了类别，还针对属性间的相关性在一定程度上弱化了属性条件独立性假设。CSAVFWNB 算法主要考虑类别与属性值频率之间存在一定的联系，同时类依赖属性值频率与类依赖属性值个数之间仍然呈正相关。根据实验结果表明，本文提出的两种类依赖实例加权算法在 10 个 UCI 数据集上皆优于 NB、IWNB、EAWIWNB 以及 AVFWNB 算法。我们提出的算法综合类别和属性取值的分布情况，但没有考虑到各个实例与测试实例的关系，从实例对分类的作用来看，如果能够综合三者进行考虑，效果应该会更好。因此，如何综合三者考虑构造实例权重，值得后期进一步研究。

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] Friedman, N., Geiger, D. and Goldszmidt, M. (1997) Bayesian Network Classifiers. *Machine Learning*, **29**, 131-163. https://doi.org/10.1023/A:1007465528199

[2] Webb, G.I., Boughton, J.R. and Wang, Z. (2005) Not So Naive Bayes: Aggregating One-Dependence Estimators. *Machine Learning*, **58**, 5-24. https://doi.org/10.1007/s10994-005-4258-6

[3] Jiang, L., Zhang, H. and Cai, Z. (2008) A Novel Bayes Model: Hidden Naive Bayes. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1361-1371. https://doi.org/10.1109/TKDE.2008.234

[4] Wu, J., Pan, S., Zhu, X., *et al*. (2016) SODE: Self-Adaptive One-Dependence Estimators for Classification. *Pattern Recognition*, **51**, 358-377. https://doi.org/10.1016/j.patcog.2015.08.023

[5] Harzevili, N.S. and Alizadeh, S.H. (2018) Mixture of Latent Multinomial Naive Bayes Classifier. *Applied Soft Computing*, **69**, 516-527. https://doi.org/10.1016/j.asoc.2018.04.020

[6] Yu, L., Jiang, L., Wang, D., *et al*. (2017) Attribute Value Weighted Average of One-Dependence Estimators. *Entropy*, **19**, 501-517. https://doi.org/10.3390/e19090501

[7] Jiang, L., Zhang, L., Li, C., *et al*. (2018) A Correlation-Based Feature Weighting Filter for Naive Bayes. *IEEE Transactions on Knowledge and Data Engineering*, **31**, 201-213. https://doi.org/10.1109/TKDE.2018.2836440

[8] Zhang, H. and Sheng, S. (2004) Learning Weighted Naive Bayes with Accurate Ranking. 4*th IEEE International Conference on Data Mining* (*ICDM*'04), Brighton, 1-4 November 2004, 567-570.

[9] Hall, M. (2006) A Decision Tree-Based Attribute Weighting Filter for Naive Bayes. *Knowledge-Based Systems*, **20**, 59-70. https://doi.org/10.1007/978-1-84628-663-6_5

[10] Jiang, L. and Zhang, H. (2006) Learning Naive Bayes for Probability Estimation by Feature Selection. *Advances in Artificial Intelligence*: 19*th Conference of the Canadian Society for Computational Studies of Intelligence*, *Canadian AI* 2006, Québec City, 7-9 June 2006, 503-514. https://doi.org/10.1007/11766247_43

[11] El Hindi, K. (2014) Fine Tuning the Naïve Bayesian Learning Algorithm. *AI Communications*, **27**, 133-141. https://doi.org/10.3233/AIC-130588

[12] Zhang, H. and Jiang, L. (2022) Fine Tuning Attribute Weighted Naive Bayes. *Neurocomputing*, **488**, 402-411.

https://doi.org/10.1016/j.neucom.2022.03.020

[13] Hindi, E.M.K., Aljulaidan, R.R. and AlSalman, H. (2020) Lazy Fine-Tuning Algorithms for Naïve Bayesian Text Classification. *Applied Soft Computing Journal*, **96**, Article ID: 106652. https://doi.org/10.1016/j.asoc.2020.106652

[14] Diab, M.D. and Hindi, E.M.K. (2016) Using Differential Evolution for Fine Tuning Naïve Bayesian Classifiers and Its Application for Text Classification. *Applied Soft Computing*, **54**, 183-199. https://doi.org/10.1016/j.asoc.2016.12.043

[15] Hindi, E.K. (2018) Combining Instance Weighting and Fine Tuning for Training Naïve Bayesian Classifiers with Scant Training Data. *The International Arab Journal of Information Technology*, **15**, 1099-1106.

[16] Xie, Z., Hsu, W., Liu, Z., *et al*. (2002) Snnb: A Selective Neighborhood Based Naive Bayes for Lazy Learning. *Advances in Knowledge Discovery and Data Mining*: 6*th Pacific-Asia Conference*, *PAKDD* 2002, Taipei, 6-8 May 2002, 104-114. https://doi.org/10.1007/3-540-47887-6_10

[17] Frank, E., Hall, M. and Pfahringer, B. (2012) Locally Weighted Naive Bayes.

[18] Jiang, L., *et al*. (2010) Improving Naive Bayes for Classification. *International Journal of Computers & Applications*, **32**, 328-332. https://doi.org/10.2316/Journal.202.2010.3.202-2747

[19] Xu, W., Jiang, L. and Yu, L. (2019) 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

[20] 杨柳, 胡桂开, 彭萍, 曾嘉琪. 嵌入属性加权的实例加权朴素贝叶斯算法[J]. 应用数学进展, 2023, 12(5): 2392-2401.