Advances in Applied Mathematics 应用数学进展, 2023, 12(5), 2392-2401
Published Online May 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.125241

Hans汉斯

嵌入属性加权的实例加权朴素贝叶斯算法

杨　柳，胡桂开*，彭　萍，曾嘉琪

东华理工大学理学院，江西　南昌

收稿日期：2023年4月28日；录用日期：2023年5月21日；发布日期：2023年5月29日

摘　要

朴素贝叶斯是一类应用广泛的分类算法，它是根据贝叶斯定理和属性条件独立来实现的。然而，属性条件独立性假设在现实生活中难以满足，为减少该假设对朴素贝叶斯算法效果的影响，本文提出了一种将属性加权嵌入到实例加权过程中的朴素贝叶斯算法。首先，基于相关性属性加权算法计算各个属性的权重；其次，将实例众数与训练实例的相似度进行属性加权，并按照不同实例众数对加权后的相似度进行算术平均得到实例权重；然后，利用实例权重构建加权朴素贝叶斯分类器；最后，采用标准UCI数据集将我们提出的算法和朴素贝叶斯算法、实例加权朴素贝叶斯算法进行仿真实验，结果表明我们提出的算法在准确率以及F1值上优于其它两种算法。

关键词

朴素贝叶斯，实例加权，属性加权

Embedded-Attribute Weighted Instance-Weighted Naive Bayes

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

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

Received: Apr. 28th, 2023; accepted: May 21st, 2023; published: May 29th, 2023

Abstract

Naive Bayes is a widely used classification algorithm, which is independently implemented based on Bayesian theorem and attribute conditions. However, the assumption of attribute conditional independence is difficult to meet in real life. To reduce the impact of this assumption on the performance of naïve Bayesian algorithms, we propose a naive Bayes algorithm by embedding attribute

*通讯作者。

文章引用: 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. DOI: 10.12677/aam.2023.125241

weighting into instance weighting process. Firstly, the weight of each attribute is calculated based on the correlation attribute weighting algorithm. Secondly, the similarity between the instance mode and the training instance is weighted by attribute, and the weighted similarity is arithmetically averaged according to the different mode instances to get the instance weight. Then, a weighted naive Bayes classifier is constructed using case weights. Finally, the standard UCI data set is used to simulate the proposed algorithm, naive Bayes algorithm and case weighted naive Bayes algorithm. The results show that the proposed algorithm is superior to the other two algorithms in accuracy and F1 value.

Keywords

Naive Bayes, Case Weighting, Attribute Weighting

1. 引言

分类是数据挖掘中一项非常重要的任务，被广泛应用疾病诊断、图像处理、垃圾邮件识别等多个领域。分类的目标是构造一个分类器，对属性集描述的实例指定合适的类标签。分类效果与数据集的规模和属性密切相关，针对不同的数据集，需要选择相应的分类算法，常见的分类算法包括朴素贝叶斯(Naive Bayes, NB)、支持向量机、决策树、KNN、神经网络等。其中，朴素贝叶斯算法是基于贝叶斯定理与属性间条件独立假设的有监督机器学习分类算法，它的分类效果以及计算复杂度可以与决策树、KNN 等媲美，并具有结构简单，高效和稳健等优点，被评为十大数据挖掘算法之一[1]。因此，近年来朴素贝叶斯分类器受到广大研究者的重视，关于朴素贝叶斯分类器的改进和应用成为一个热点。朴素贝叶斯算法的核心思想是通过考虑属性概率来预测分类，即对于给定的测试样本，求解在此样本出现条件下各个类别出现的概率，将测试样本归于发生概率最大的那一类。假设 x 为包含 m 个属性的测试实例，记 $x=(a_1,a_2,\cdots,a_m)$，则 x 的类标签可通过公式(1)获得。

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

其中 C 是所有类组成的集合，c 是 C 中的一个类向量，m 为属性的个数，a_j 为测试实例 x 的第 j 个属性的属性值，$P(c)$ 是各类别的先验概率，$P(a_j|c)$ 为类别 c 条件下 a_j 发生的条件概率。为避免概率为零，公式(1)中的先验概率 $P(c)$ 和条件概率 $P(a_j|c)$ 分别利用了拉普拉斯平滑方法计算，见以下公式，

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

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

其中 n 为训练实例的个数，c_i 为第 i 个训练实例的类标签，k 为 C 中种类个数，a_{ij} 为第 i 个训练实例中第

j 个属性的属性值，n_j 为第 j 个属性对应不同属性值的个数，$\delta(a,b)$ 是示性函数，当 $a=b$ 时，函数值为1，当 $a\neq b$ 时，函数值为0。

朴素贝叶斯算法因要求属性之间条件独立而得名，然而各个属性之间相互独立在现实生活中往往很难实现，在这种情况下盲目使用朴素贝叶斯分类算法不仅影响算法的准确率，也影响算法的可信度。因此，如何减小条件独立性对朴素贝叶斯算法的负面影响成为算法改进的主要方向，目前概括为四个方向：结构扩展、微调、属性加权(属性选择)和实例加权(实例选择)。

结构扩展是对具有相关性的属性之间添加有向边，每个属性之间添加有向边后形成一个有向无环图，清晰地展示了属性之间的网络拓扑结构，从而减轻属性之间条件独立性假设的限制，提高分类准确率。文献中，常用的结构扩展算法主要有以下三种：如，Friedman 等[2]提出的树扩展朴素贝叶斯，其网络拓扑结构如树权结构，将类变量作为树结构的根节点，所有属性中确定一个除类变量根节点以外的节点作为属性根节点，通过计算属性间的互信息来建立有向边，沿着属性根节点确定有向边的方向建立网络结构；Webb 等学者[3]提出的平均单依赖估测器，该模型可以看成是一个特殊的树扩展朴素贝叶斯分类器，除了类根节点外，所有属性都有机会成为属性根节点，即树扩展结构图的个数由属性个数决定，最后朴素贝叶斯模型的建立将综合考虑所有树扩展结构图的结果；Jiang 等学者[4]提出了隐朴素贝叶斯，该算法为每一个目标属性生成一个隐藏的父亲结点，该父亲结点综合了其他所有属性对目标属性的影响，避免学习拓扑结构。目前，也有很多新的结构扩展朴素贝叶斯算法，见文献[5] [6] [7]。微调法是在朴素贝叶斯分类的基础上对错判后的条件概率进行调整，然后再进一步构建分类器的改进方法，其核心要点是条件概率的修正，具体可参阅文献[8] [9] [10]。

在朴素贝叶斯分类器中，要求每个属性同等重要。然而现实应用中，不同属性在分类判别中承担着不同的角色，部分属性的作用高于其它属性。自然地，改进朴素贝叶斯方法就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。属性加权朴素贝叶斯模型是建立在加权后的数据上，如何确定各个属性的权重是改进朴素贝叶斯算法的关键问题，属性加权方法可以分为过滤法和包装法两种[11]。过滤法在构建朴素贝叶斯分类器之前就直接对训练集的数据特性进行评估，并利用数据直接计算出属性的权值，然后利用加权后的数据来训练模型；而包装法将算法学习包裹在属性加权过程中，通过优化权值，层层迭代来提升朴素贝叶斯的分类性能。常用的过滤法属性加权主要有以下几种。如，Zhang 和 Sheng [12]提出了基于增益率的属性加权方法，该法认为收益率高的属性应该分配更大的权重；Hall [13]提出了基于决策树深度的属性加权方法，该方法通过构建未修剪的决策树和查看树中目标属性的深度来估计属性依赖的程度；Lee 等学者[14]利用 Kullback-Leibler 散度来计算属性权重；Jiang 等学者[15]提出了一种基于相关性的属性加权滤波器(Correlation-based Feature Weighting, CFW)，认为高预测能力属性应该与类高度相关，即有最大相关性，而与其余属性不相关，即有着最小冗余度，由此给出了属性权重的计算公式。关于包装法属性加权，学者们也提出了多种方法，如基于相关性的特征选择[16]、基于决策树的属性过滤算法[17]、基于条件似然对数的选择性朴素贝叶斯算法[18]等。

实例加权(实例选择)与属性加权(属性选择)类似，只是在实例的维度上处理数据。实例选择也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。与属性加权一样，不仅不同属性对类的影响不一样，不同实例对类的作用也不一样。但和属性加权不一样的是实例与类之间的相关性不好衡量，主要通过衡量训练实例与测试实例间的距离来确定权重。因此，该方面的研究文献不多。如，Xie 等[19]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[20]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；Jiang 等[21]提出了一种实例加权朴素贝叶斯算法(Instance Weighted Naive Bayes, IWNB)，该文献首次提出了属性值众数，实例众数的概念，并通过计算每个实例和实例众数之间的相似度来确定实例权重；Xu 等[22]提出了属性值频率加权朴素贝叶斯，该算法通过计算属性值频率向量与属性所有值个数向量的内积得到实例权重。

IWNB 算法思想简单，在各个数据集的应用中亦取得不错的效果。然而，IWNB 算法在计算实例众数时，没有考虑到每个属性可能会有多个属性值众数的情况，从而有多个实例众数现象出现，只考虑一个实例众数就不够全面。另外，在计算实例权重时，只是简单的对相似度进行属性求和，正如上文所述，每个属性对分类的影响不一样，故考虑将属性权重嵌入进去，进行属性加权求和。为此，本文在综合考虑多个实例众数和属性对分类影响程度的基础上拟提出一种新的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)。

下文结构安排如下：第 2 节介绍 CFW 和 IWNB 等相关工作；第 3 节提出 EAWIWNB 算法；第 4 节给出 EAWIWNB、NB 和 IWNB 三个算法在不同数据集的实验对比；第 5 节为结论总结和展望。

2. 相关工作

2.1. CFW 属性加权

属性加权是给不同重要程度的属性分配相应的权重，减缓条件独立性的限制，从而提高算法整体的分类效率。过滤法的步骤是先计算各个属性的权重，再利用加权后的数据建立朴素贝叶斯模型，此时朴素贝叶斯模型中先验概率 $P(c)$ 和条件概率 $P(a_v|c)$ 的计算公式不变，而属性加权后的朴素贝叶斯公式则变为：

$$c(x)=\arg\max_{c\in C} P(c)\prod_{v=1}^{m}P(a_v|c)^{W_v}, \tag{4}$$

其中，W_v 是第 v 个属性的属性权重。

CFW 算法认为与类具有高度相关性(相关性)，并且与其余属性不具有相关性(冗余度)的属性应赋予其更高的权重[15]，并利用互信息来度量相关性与冗余度的大小，具体公式如下：

$$I(A_v;C)=\sum_{a_v}\sum_{c}P(a_v,c)\log\frac{P(a_v,c)}{P(a_v)P(c)}, \tag{5}$$

$$I(A_v;A_u)=\sum_{a_v}\sum_{a_u}P(a_v,a_u)\log\frac{P(a_v,a_u)}{P(a_v)P(a_u)}, \tag{6}$$

其中 A_v 和 A_u 为两个不同的属性，A_v 为目标属性。a_v 和 a_u 是对应属性的属性值，C 为类向量，c 为类向量中的类。为了方便计算，对上述计算的互信息进行归一化：

$$NI(A_v;C)=\frac{I(A_v;C)}{\frac{1}{m}\sum_{v=1}^{m}I(A_v;C)}, \tag{7}$$

$$NI(A_v;A_u)=\frac{I(A_v;A_u)}{\frac{1}{m(m-1)}\sum_{v=1}^{m}\sum_{u=1\Lambda u\neq v}^{m}I(A_v;A_u)}, \tag{8}$$

其中 m 为属性个数。在确定权重时既要使属性与类的相关性 $NI(A_v;C)$ 高，也要使属性与其余属性的冗

余度 $NI(A_v;A_u)$ 低，因此，目标属性 A_v 的权重应该与 $NI(A_v;C)$ 和平均冗余度的差成正比：

$$D_v = NI(A_v;C) - \frac{1}{m-1}\sum_{u=1\Lambda u\neq v}^{m} NI(A_v;A_u). \tag{9}$$

公式(9)计算出的属性权重有可能为负数，为了符合实际情况，利用 Sigmoid 函数——对数几率函数将权重调整为 0 到 1 之间，最终目标属性 A_v 权重计算公式定义为：

$$W_v = \frac{1}{1+e^{-D_v}}. \tag{10}$$

计算出所有属性权重后，再利用公式(4)构建朴素贝叶斯分类器。

2.2. IWNB 实例加权

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，同样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：

$$c(x) = \arg\max_{c\in C} P(c)\prod_{j=1}^{m} P(a_j|c) \tag{11}$$

在 IWNB 分类器中，目标函数公式与朴素贝叶斯公式(1)相同，但 $P(c)$ 和 $P(a_j|c)$ 的计算方式不同，在求和过程中考虑了实例加权，具体公式如下：

$$P(c) = \frac{\sum_{i=1}^{n} w_i\delta(c_i,c)+1}{\sum_{i=1}^{n} w_i + k}, \tag{12}$$

$$P(a_j|c) = \frac{\sum_{i=1}^{n} w_i\delta(a_{ij},a_j)\delta(c_i,c)+1}{\sum_{i=1}^{n} w_i\delta(c_i,c)+n_j}, \tag{13}$$

其中 w_i 是第 i 个训练实例的权重。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每个实例的权重，参照文献[19]给出属性值众数、实例众数以及相似度的定义如下：

定义 1：属性值众数是属性的所有值中出现频率最高的属性值。

定义 2：实例众数是所有属性值都由属性值众数构成的实例。

定义 3：实例 x 与实例 y 之间的相似度定义为：

$$s(x,y) = \sum_{j=1}^{m}\delta(a_j(x),a_j(y)), \tag{14}$$

其中 m 为属性个数，$a_j(x)$ 与 $a_j(y)$ 分别表示实例 x 和实例 y 的第 j 个属性值。

IWNB 算法首先通过训练数据计算出实例众数 z，其次计算每个实例 x 与 z 的相似度，然后给每个实例 x 赋予权重 $1+s(x,z)$，其中加 1 是为了防止权重等于 0，最后利用加权后的数据建立朴素贝叶斯模型。其算法流程如表 1 所示。

虽然 IWNB 算法取得了不错的效果，但该算法没有考虑到出现多个实例众数的情形，同时在进行相似度求和过程，没有区分属性间的差异。为解决以上问题，我们提出 EAWIWNB 算法对 IWNB 算法进行改进，具体过程见下节。

Table 1. IWNB algorithmic flow
表 1. IWNB 算法流程

算法：IWNB 算法
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：计算训练实例集 D 的实例众数 z
Step2：利用公式(13)计算每个训练实例 y 与 z 的相似度 $s(y,z)$
Step3：给每个训练实例赋予权重 $1+s(y,z)$
Step4：利用加权后的 D 建立朴素贝叶斯模型
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

3. 嵌入属性加权的实例加权朴素贝叶斯

在 IWNB 算法实施过程，我们发现部分属性中出现频率最高的属性值并不唯一，即存在多个属性值众数的情况，从而导致实例众数不唯一，若单纯用一个实例众数计算相似度，容易出现偏差。考虑到这种情况，EAWIWNB 算法将所有属性值众数进行组合得到多个实例众数，分别计算训练实例与每个实例众数的相似度，再做算术平均。首先给出多个实例众数的举例说明：假设数据有 3 个属性 a_1、a_2、a_3，其中 a_1 的属性值众数有 a_{11}、a_{12}、a_{13}；a_2 的属性值众数有 a_{21}；a_3 的属性值众数有 a_{31}、a_{32}。所有属性值众数的组合情况相加最终将会得到 $C_3^1\times C_1^1\times C_2^1$ 个实例众数。即通过组合可以得到 6 个实例众数，具体如图 1 所示。

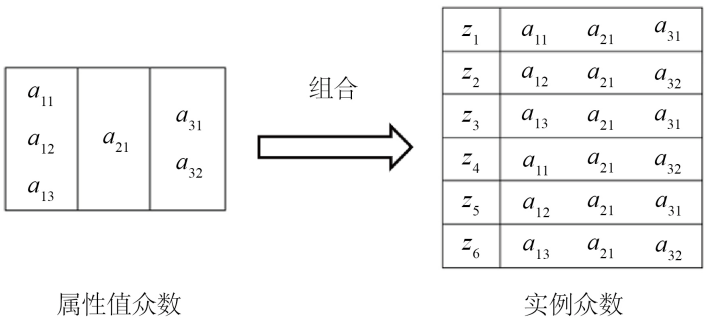

Figure 1. Construction of instance mode
图 1. 实例众数的构建

另外，IWNB 算法在利用公式(13)计算相似度时，没有考虑到每个属性对分类的影响程度，只是简单的算术平均累加。然而，正如 CFW 属性加权算法中提到，高预测能力属性应该与类高度相关，有最大相关性，而与其余属性不相关，有着最小冗余度，为此我们拟在公式(13)中嵌入属性权重。EAWIWNB 算法综合了所有实例众数的信息，同时考虑了属性间的重要程度，在实例加权的同时嵌入了属性加权，是实例加权和属性加权的融合，比单独考虑其中一个方面能够更好地减轻属性之间条件独立性带来的影响。近来，Zhang [1]提出了将属性加权与实例加权相结合的分类方法，但其算法步骤是先由实例加权调整条件概率，再利用属性加权调整目标函数，是二者的前后混合，而我们提出的算法是在实例加权过程中嵌

入融合。实例权重计算步骤如下：

1) 利用 2.1 节中的 CFW 算法计算训练数据中每个属性的权重，记为$W_j, j=1,2,\cdots,m$。

2) 利用 2.2 节中的定义 1-2 计算训练数据中的所有实例众数。假设训练数据有 T 个实例众数，不妨记为 $z=(z_1,\cdots,z_T)$。

3) 改进公式(13)计算实例 x 与每个众数实例 z_t 的相似度，计算公式为：

$$s'(x,z_t)=\sum_{j=1}^{m}W_j\delta(x_j,z_{tj}),\ t=1,2,\cdots,T, \tag{15}$$

其中 T 表示实例众数的个数，m 表示属性个数，x_j 表示实例 x 的第 j 个属性值，z_{tj} 表示第 t 个实例众数的第 j 个属性值。

4) 根据公式 $s(x,z)=\frac{1}{T}\sum_{t=1}^{T}s'(x,z_t)$ 计算实例 x 与众数实例的相似度，最后得到每个实例 x 的权重为 $1+s(x,z)$，即

$$w_i=1+\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m}W_j\delta(x_j,z_{tj}) \tag{16}$$

最后用嵌入了属性权重的实例加权数据建立朴素贝叶斯模型。EAWIWNB 的所有算法流程见表 2。

Table 2. EAWIWNB algorithmic flow
表 2. EAWIWNB 算法流程

算法：EAWIWNB 算法
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：利用 CFW 算法计算 D 的所有属性权重 w_j
Step2：计算训练实例集 D 的所有实例众数 z
Step3：利用公式(14)计算训练实例与每个众数实例的相似度
Step4：利用公式(15)计算每个训练实例的权重
Step5：利用实例加权思想建立朴素贝叶斯模型
Step6：利用模型预测测试实例 x_{test} 的类标签
Step7：输出 x_{test} 的类标签

4. 实验结果与分析

4.1. 数据集与实验环境

为了验证 EAWIWNB 算法的分类效率，本节筛选了 6 个 UCI 数据集(如表 3 所示)进行算法实践，这些数据集包含医疗和社会生活等多个方面，能有效验证算法的应用性，并且它们都属于离散型数据，其中 Breast cancer、Dermatology、Mammographic 三个数据集含有少量的缺失值，因缺少数据占比不大，我们直接将含有缺失值的实例样本删除，便于后期算法运行。

实验环境：本文所有实验都是在系统 win10，512G (SSD)硬盘，i5-1135G7 的 CPU，内存 16G 的 PC 机上通过 R 4.2.2 版本完成。

Table 3. Data set information
表 3. 数据集信息

数据集	实例数	特征数	类别数	是否含有缺失值
Hayes roth	160	5	3	否
Breast cancer	286	9	2	是
Dermatology	366	33	6	是
Lymphography	148	18	4	否
Mammographic	961	6	2	是
Somerville happiness survey	143	7	2	否

4.2. 算法验证与分析

接下来，我们利用预处理后的 6 个数据集对 EAWIWNB 算法和 IWNB 算法、NB 算法进行实验比较。首先，采用交叉验证把数据随机分成十份，80%作为训练集，20%作为测试集；其次，分别利用训练集学习三个不同的分类方法，并对测试集进行检验，统计各个算法的准确率；最后，依次重复上述步骤 10 遍，分别计算不同数据集在不同算法下准确率的均值及标准差，并采用 t 检验在显著性水平 0.05 下分别判断 EAWIWNB 算法与 NB 算法、EAWIWNB 算法与 IWNB 算法是否存在显著性差异，若有显著差异，则用符号“√”表示，无显著差异，则用符号“○”表示，计算结果见表 4。

Table 4. Accuracy comparison
表 4. 准确率对比

数据集	准确率均值		
	NB	IWNB	EAWIWNB
Hayes roth	0.8 ± 0.07√	0.84 ± 0.08√	0.86 ± 0.06
Breast cancer	0.75 ± 0.03√	0.76 ± 0.04√	0.78 ± 0.03
Dermatology	0.97 ± 0.01○	0.96 ± 0.02√	0.96 ± 0.02
Lymphography	0.86 ± 0.06○	0.87 ± 0.06○	0.87 ± 0.06
Mammographic	0.82 ± 0.03√	0.82 ± 0.03√	0.83 ± 0.03
Somerville happiness survey	0.61 ± 0.08√	0.61 ± 0.07√	0.64 ± 0.07

从表 4 可以看出，在数据集 Dermatology、Lymphography 中算法 EAWIWNB 和 NB 无显著差异，但在数据集 Hayes roth、Breast cancer、Mammographic、Somerville happiness survey 中有显著差异，并且我们提出的 EAWIWNB 算法在平均准确率上优于 NB 算法；而对于算法 IWNB 和 EAWIWNB 的比较，二者仅在数据集 Lymphography 上无显著差异，在其余所有数据中都存在显著差异，我们提出的 EAWIWNB 算法在平均准确率上几乎全部大于 IWNB 算法。因此，算法 EAWIWNB 在准确率上明显优于 NB 及 IWNB，即 EAWIWNB 算法的改进显著的提高了朴素贝叶斯算法的分类准确率。

此外，为了使实验结果更有说服力，我们还将三种算法的 F1 值进行了比较。F1 值综合了精确度和召回率的信息，也是衡量算法精度的主要指标。该指标数值越大说明模型效果越好。对于二元分类数据，我们直接计算每个算法的平均值；对于多元分类数据，我们将其划分为多个二元分类，分别计算每个二元分类的 F1 值，并取平均值。具体结果见表 5。

Table 5. F1 value comparison
表 5. F1 值对比

数据集	F1 值		
	NB	IWNB	EAWIWNB
Hayes roth	0.7916667	0.7916667	0.8100665
Breast cancer	0.8539326	0.8636364	0.8666667
Dermatology	0.9505495	0.9505495	0.9629630
Lymphography	0.8036636	0.8944193	0.9181287
Mammographic	0.8157895	0.8157895	0.8387097
Somerville happiness survey	0.6206897	0.5925926	0.6666667

从表 5 可以看出，EAWIWNB 算法在 F1 值上明显优于其余两种算法，说明 EAWIWNB 在精确度以及召回率上也有显著优势。综合以上两方面的比较可知，EAWIWNB 算法的表现分别优于 NB 和 IWNB 算法，改进有效。

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] Zhang, H., Jiang, L.X. and Yu, L.J. (2021) Attribute and Instance Weighted Naive Bayes. *Pattern Recognition*, **111**, 107674-107684. https://doi.org/10.1016/j.patcog.2020.107674

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

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

[4] Jiang, L.X., Zhang, H. and Cai, Z.H. (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

[5] Wu, J., Pan, S.R., Zhu, X.Q., *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

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

[7] 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

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

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

[10] Zhang, H. and Jiang, L.X. (2022) Fine Tuning Attribute Weighted Naive Bayes. *Neurocomputing*, **488**, 402-411. https://doi.org/10.1016/j.neucom.2022.03.020

[11] 余良俊. 属性加权的贝叶斯网络分类算法及其应用研究[D]: [博士学位论文]. 武汉: 中国地质大学, 2018.

[12] Zhang, H. and Sheng, S.L. (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.

[13] Hall, M. (2007) A Decision Tree-Based Attribute Weighting Filter for Naive Bayes. In: *Research and Development in Intelligent Systems XXIII*: *Proceedings of AI*-2006, *the* 26*th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, Springer, London, 59-70. https://doi.org/10.1007/978-1-84628-663-6_5

[14] Lee, C.H., Gutierrez, F. and Dou, D. (2011) Calculating Feature Weights in Naive Bayes with Kullback-Leibler Meas-

ure. 2011 *IEEE* 11*th International Conference on Data Mining*, Vancouver, 11 December 2011, 1146-1151. https://doi.org/10.1109/ICDM.2011.29

[15] Jiang, L.X., Zhang, L.G., Li, C.Q., *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

[16] Hall, M.A. (2000) Correlation-Based Feature Selection of Discrete and Numeric Class Machine Learning. 2000 Working Papers, 5: 00/08.

[17] Ratanamahatana, C. and Gunopulos, D. (2003) Feature Selection for the Naive Bayesian Classifier Using Decision Trees. *Applied Artificial Intelligence*, **17**, 475-487. https://doi.org/10.1080/713827175

[18] Jiang, L.X. 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

[19] Xie, Z.P., Hsu, W., Liu, Z.T., *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

[20] Frank, E., Hall, M. and Pfahringer, B. (2003) Locally Weighted Naive Bayes. *The Conference on Uncertainty in Artificial Intelligence*, Acapulco, 7-10 August 2003, 249-256.

[21] Jiang, L.X., Cai, Z.H. and Wang, D.H. (2010) Improving Naive Bayes for Classification. *International Journal of Computers and Applications*, **32**, 328-332. https://doi.org/10.2316/Journal.202.2010.3.202-2747

[22] Xu, W.Q., Jiang, L.X. and Yu, L.J. (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