Advances in Applied Mathematics 应用数学进展, 2020, 9(5), 752-758
Published Online May 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.95089

Hans汉斯

A Weighting Attribute Method for Classification Problems

Tingfang Wang, Dongxi Li

School of Mathematics, Taiyuan University of Technology, Jinzhong Shanxi
Email: tingfangwang@163.com

Received: May 1st, 2020; accepted: May 14th, 2020; published: May 21st, 2020

Abstract

Attribute weighting adjustments are used in machine learning models to improve performance. In this paper, we propose a novel attribute weighting method based on mutual information and apply this method to two classical machine learning models for classification. We study the performance of our weighting method by conducting experiments on the Wisconsin Breast Cancer database. In both machine learning models, our weighted attribute models tend to outperform the corresponding conventional machine learning models in classification which also approves that our weighting method is reasonable and applicable.

Keywords

Weighted Attribute, Classification Problem, Naïve Bayes, *k*-Nearest Neighbor

一种用于分类问题的属性加权方法

王庭芳，李东喜

太原理工大学数学学院，山西 晋中
Email: tingfangwang@163.com

收稿日期：2020年5月1日；录用日期：2020年5月14日；发布日期：2020年5月21日

摘　要

属性加权调整通常用于机器学习方法中以提高这些方法的性能。在本文中，我们提出了一种基于互信息的新型属性加权方法，并将该方法应用于两种经典的机器学习分类方法中。我们通过在威斯康星州乳腺癌数据集进行实验来研究加权方法的性能。我们的实验结果表明，针对分类任务，我们的加权机器学习方法往往优于相应的传统机器学习方法，从而证明了本文提出的加权方法的合理性和实用性。

文章引用: 王庭芳, 李东喜. 一种用于分类问题的属性加权方法[J]. 应用数学进展, 2020, 9(5): 752-758.
DOI: 10.12677/aam.2020.95089

关键词

属性加权，分类问题，朴素贝叶斯，*k*近邻方法

1. 引言

分类问题是监督学习的一个核心问题。在监督学习中，当输出变量Y取有限个离散值时，预测问题就变成了分类问题[1]。任何分类算法的目标都是基于具有类别的给定数据构建模型，然后使用该模型对新实例进行分类。对分类方法的研究主要用于解决实际问题，获得至少比猜测更好的分类精度。例如，在乳腺癌测试中，广泛使用的乳腺癌诊断工具是细针穿刺细胞学检查(Fine Needle Aspiration Cytology: FNAC)。FNAC 的正确分类率约为 90%，因此迫切需要改进分类系统，协助医生进行诊断来提高乳腺癌诊断的准确率[2]。

在常规的分类算法中，通常假定属性对类标签具有完全相同的影响，并且在不考虑特征对类的可能存在不同程度影响的情况下构建分类模型。然而，在大多数现实情况下，不同属性对类标签确实存在不同程度的影响。加权分类可以有效地解决此类问题，通过为不同属性赋予相应的权重，以区分其在分类任务中不同的重要程度来获得更好的分类模型。将属性加权方法应用到机器学习模型中的研究越来越受到关注。Zaidi 等人在 2013 年的研究中提出一种加权朴素贝叶斯算法，该算法以最小化均方误差函数为目标，通过赋予分类特征权重来弱化条件独立性假设[3]；Xueson Yan 等人在 2016 年的相关研究中提出了一种双重加权朴素贝叶斯分类器，将属性加权与样本加权共同整合到分类模型中，用于多分类任务[4]。1977 年，Wettschereck 和 Mohri 提出了一种区分懒惰学习算法中权重设置的框架，并通过实验比较分析了各个方法的性能[5]。Gipta 在 2012 年的研究中提出一种动态加权 k 近邻分类算法，该算法通过 k-Means 聚类方法来确定属性权重[6]。

在本文中，我们提出了一种基于互信息的权重确定方法，并将该方法应用到两种经典的机器学习方法中，提出加权朴素贝叶斯分类器和加权 k 近邻方法。以解决分类任务中的属性加权问题。我们通过将这些方法的性能与传统方法的性能进行比较来评估本文提出的加权分类方法。通过实验验证本文提出的基于互信息的加权方法的有效性和合理性。

本文的其余部分安排如下。第 2 部分介绍相关理论知识和方法。第 3 部分介绍了本文的实验和相应的结果。最后，第 4 部分对本文提出的方法进行了分析讨论。

2. 方法

2.1. 权重获取

在统计学中，权函数是指在计算平均数等指标时，为各个变量值计算轻重作用的数值的函数。权函数可以被视为执行求和、积分、乘积和平均这些运算的数学系统[2]。它的主要作用是增强同一数据集中某些变量对结果的影响。在统计分析中，经常会用到权函数的概念，如加权和以及加权平均数。根据取值的不同，权重可以分为离散型和连续型权重，不同的取值类型有不同的定义方式。本文仅讨论离散条件下的权函数。

权函数$\omega: X \to \mathbb{R}^+$定义为离散集$X$上的取值为正的函数。本文使用的加权和以及加权积分别定义为：

$$\sum_{x\in X} f(x)\omega(x) \tag{1}$$

$$\prod_{x\in X} f(x)\omega(x) \tag{2}$$

对于分类任务，特征加权是提高分类准确性的有效技术，尤其是对于那些不同特征对类别标签有不同程度影响的数据集。由于互信息是两个变量之间相互独立性的度量，互信息越大，说明其中一个变量中所包含的有关另一个变量的信息越多，从而表明两个变量之间的联系越紧密。并且，标准化互信息$[0,1]$的取值范围与权重的取值范围契合，节省了专门进行权重归一化的时间。因此我们将每个特征与类标签之间的标准化互信息作为其权重。假设数据集T的样本量为N，X为数据集中的一个离散型特征，C为类标签变量。其中X、C的取值分别为$X=\{x_1,x_2,\cdots,x_n\}$和$C=\{c_1,c_2,\cdots,c_K\}$，则属性X的每个取值的权重$\omega(x_i)$定义如下：

$$\omega(x_i)=NMI(X,C)=\frac{MI(X,C)}{mean(H(X),H(C))} \tag{3}$$

其中，

$$MI(X,C)=\sum_{i=1}^{n}\sum_{j=1}^{K}\frac{|x_i\cap c_j|}{N}\log\left(\frac{N|x_i\cap c_j|}{|x_i||c_j|}\right) \tag{4}$$

$$H(X)=-\sum_{i=1}^{n}P\left(\frac{|x_i|}{N}\right)\log\left(\frac{|x_i|}{N}\right) \tag{5}$$

$$H(C)=-\sum_{i=1}^{K}P\left(\frac{|c_i|}{N}\right)\log\left(\frac{|c_i|}{N}\right) \tag{6}$$

2.2. 加权朴素贝叶斯分类器

朴素贝叶斯分类器是贝叶斯决策论(Bayesian decision theory)中的一种基本分类方法。在所有相关概率都已知的理想情况下，贝叶斯决策论考虑如何基于这些概率和误判损失来选择最优的类别标记[7]。在朴素贝叶斯分类器中，由于条件独立性假设的存在，使得所有分类特征对于计算后验概率的贡献相同，然而，在大多数实际应用中，情况并非如此。为了解决这一问题，研究者们提出了一些半朴素贝叶斯分类器，即在分类过程中适当考虑某些属性之间的相互依赖关系，如TAN (Tree Augmented naïve Bayes)以及AODE (Averaged One-Dependent Estimator)。

本文提出加权朴素贝叶斯分类器通过在计算后验概率时对属性加权来完成分类任务，以减少属性条件独立性假设带来的消极影响。假设输入空间$\mathcal{X}\subseteq\mathbb{R}^n$是$n$维向量的集合，输出空间为类别标签集合$\mathcal{Y}=\{C_1,C_2,\cdots,C_K\}$. 用$X=\left\{X^{(1)},X^{(2)},\cdots,X^{(n)}\right\}$表示输入空间$\mathcal{X}$上的随机向量，其中$X^{(i)},i=1,2,\cdots,n\subseteq\mathbb{R}^n$表示数据集中的第$i$个特征，用$Y$表示输出空间$\mathcal{Y}$上的随机变量。算法的各参数具体定义如下：

算法输入：数据集：$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$；待分类实例：$x=\left(x^{(1)},x^{(2)},\cdots,x^{(n)}\right)^{\mathrm{T}}$；

算法输出：类标签预测结果：$y\in\mathcal{Y}$

1) 根据数据集信息计算各属性的权重，为待分类实例中的每一个属性值分配相应的权重值：

$$\omega\left(x^{(i)}\right)=\omega\left(X^{(i)}\right)=NMI\left(X^{(i)},\mathcal{Y}\right) \tag{7}$$

2) 计算在特定类别下各属性值的加权条件概率：

$$P_{\omega}\left(x^{(i)} \mid C_k\right)=\omega\left(x^{(i)}\right)P\left(x^{(i)} \mid C_k\right) \tag{8}$$

3) 根据贝叶斯判定准则，对待分类实例的类标签进行预测：

$$y=\arg\max_{C_k} P\left(Y=C_k\right)\prod_{i=1}^{n}P_{\omega}\left(x^{(i)} \mid C_k\right) \tag{9}$$

其中，$x^{(i)}$ 为待分类实例中属性 $X^{(i)}$ 的取值。

在实际任务中，给定训练集，可将加权朴素贝叶斯分类器中涉及的各权重值以及所有概率估计值计算出来并存储在模型中，待到需要对实例进行预测时，只需要进行简单的计算便能得到分类结果，从而大大提高预测速度。

2.3. 加权 k 近邻方法

k 近邻方法是一种常用的监督学习方法。k 近邻方法是懒惰学习算法中的一种，只有在输入待分类实例后，算法才会开始运行，训练时间开销为零。算法通过距离函数来计算两两实例间的距离，从而确定待分类实例的 k 个最近邻样本点，最后根据决策规则来判断待分类实例的类标签。k 的取值不同，分类结果会有显著不同；其次，采用不同的距离计算方式，找出的近邻也可能存在显著差异，从而导致分类结果有显著不同[7]。本文采用欧氏距离作为实例点之间的距离度量，采用多数表决规则来确定待分类实例的类别标签。

本文提出的加权 k 近邻方法将加权和引入欧氏距离函数的计算中，通过

$$d_{\omega}\left(x_i,x_j\right)=\left(\sum_{l=1}^{n}\omega\left(x_i^{(l)}\right)\left|x_i^{(l)}-x_j^{(l)}\right|^2\right)^{\frac{1}{2}} \tag{10}$$

来计算实例 $x_i=\left(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)}\right)^{\mathrm{T}}$ 与实例 $x_j=\left(x_j^{(1)},x_j^{(2)},\cdots,x_j^{(n)}\right)^{\mathrm{T}}$ 之间的距离：

其中，

$$\omega\left(x_i^{(l)}\right)=NMI\left(X^{(l)},C\right) \tag{11}$$

算法输入：待分类的实例 $x=\left\{x^{(1)},x^{(2)},\cdots,x^{(n)}\right\}$；数据集 $T=\left\{\left(x_1,y_1\right),\left(x_2,y_2\right),\cdots,\left(x_N,y_N\right)\right\}$

算法输出：实例 x 的类别 y。

1) 通过加权欧氏距离 $d_W(\,)$ 的计算结果找出训练集 T 中与实例 x 最近邻的 k 个样本点，确定这 k 个点组成的邻域 $N_{\omega k}(x)$；

2) 在 $N_{\omega k}(x)$ 中根据多数表决规则决定待分类实例 x 的类别标签 y：

$$y=\arg\max_{C_j}\sum_{x_j\in N_{\omega k}(x)}I\left(y_i=C_j\right),\ i=1,2,\cdots,N;\ \ j=1,2,\cdots,K \tag{12}$$

其中，$I\left(y_i=C_j\right)$ 为指示函数，条件成立时函数取值为 1，反之取值为 0。

加权 k 近邻方法简单直观，在确定 k 值、距离度量和分类决策规则后，待分类实例的预测类标签唯一确定。

3. 实验及结果

3.1. 数据集及实验步骤

本文选用威斯康星乳腺癌数据集[8] (Wisconsin Breast Cancer Database: WBCD)进行实验，该数据集总共有 699 个实例，除类别标签外还包含其他 9 个特征属性，列举在表 1 中。根据类标签的不同，该数据集可以划分为 241 个恶性肿瘤实例和 458 个良性肿瘤实例。除此之外，该数据集还包含 16 个有缺失值的

样本，在进行实验之前，本文首先将这 16 个实例删除，实验仅采用剩余的 683 个完整实例，包括 239 个恶性肿瘤样本和 444 个良性肿瘤样本。具体实验步骤如下：

1) 首先采用传统的机器学习方法：朴素贝叶斯分类器、k 近邻方法，对数据集进行分类，得到相应算法的分类准确率。

2) 其次，采用本文提出的各类加权机器学习方法：加权朴素贝叶斯分类器、加权 k 近邻方法再次进行分类，同样以分类准确率作为评价指标。

3) 最后，通过比较这两种传统机器学习方法与其相应的加权方法在分类准确率方面的差异来对本文提出的加权机器学习方法的性能进行评估并借此验证本文加权方法的实用性。

Table 1. Features description in Wisconsin breast cancer database
表 1. 威斯康星乳腺癌数据集属性描述

数目	属性名称	取值类型	有无缺失值
1	Clump thickness	离散型	无
2	Uniformity of cell size	离散型	无
3	Uniformity of cell shape	离散型	无
4	Marginal adhesion	离散型	无
5	Single epithelial cell size	离散型	无
6	Bare nuclei	离散型	有
7	Bland chromatin	离散型	无
8	Normal nucleoli	离散型	无
9	Mitoses	离散型	无

3.2. 实验结果

对于朴素贝叶斯分类器以及本文提出的加权朴素贝叶斯，采用不同折数的交叉验证来进行实验，并计算了各折交叉验证下的分类精度以及平均精度来进行算法性能评估。例如，五折交叉验证首先将数据集划分为 5 个子集，依次选择 5 个子集中的一个作为测试集，其他 4 个子集组成的合集作为训练集。因此，作为测试集的每折数据都有一个分类准确度，这些准确度的平均值作为相应交叉验证下的分类精度。实验结果统计在表 2 中。

Table 2. Classification accuracies (%) for Naïve Bayes
表 2. 朴素贝叶斯分类器的分类准确率

交叉验证折数	朴素贝叶斯分类器	加权朴素贝叶斯分类器
3cv	97.00	97.00
5cv	97.15	97.15
7cv	97.15	97.15
9cv	97.16	97.16
10cv	97.15	97.15
11cv	97.16	97.16
Average	97.13	97.13

从实验结果可以看出，本文提出的加权朴素贝叶斯模型在分类准确率方面的表现与传统的朴素贝叶斯分类器相当，平均分类精度都为97.13%。尽管朴素贝叶斯分类器中的条件独立性假设在很多现实数据集中都不成立，但其在实际应用中的表现极佳，是机器学习方法中最有力的分类工具之一。本文提出的方法虽然没有提高其分类精度，但从实验结果也可以看出本文基于互信息的加权方法并不会对本来稳健的分类器造成消极影响，因此本文的加权方法对此类机器学习方法具有一定普适性。

对于 k 近邻方法以及相应的加权 k 近邻方法，由于其作为一种惰性学习算法的特性，不存在训练模型，从而不适用于交叉验证。k 近邻方法会事先存储整个数据集，待到有待分类实例输入后才会开始进行分类任务。代替使用交叉验证，本文采用不同的 k 值来进行相关实验，每个 k 值都对应一个分类准确度，同时本文还计算了平均分类精度。具体实验结果见于表3。

Table 3. Classification accuracies (%) for k-NN
表 3. k 近邻方法的分类准确率

k	k 近邻方法	加权 k 近邻方法
1	88.29	91.59
3	89.49	92.49
5	90.39	92.49
7	90.39	92.19
9	89.79	91.89
Average	89.67	92.13

从上述实验结果可以看出，对于每一个不同的 k 值，本文提出的加权方法对分类准确率都有提高，加权模型的最高分类精度达到92.49%，比相应 k 值下的传统模型最高提升了3%的准确率。此外，传统 k 近邻方法的平均分类精度为 89.67%，而本文提出的加权模型的平均分类精度达到 92.13%，平均提高了2.46%的分类精度。因此，可以得出本文提出的加权 k 近邻模型在分类准确率方面的表现优于传统方法这一结论，从而也可以证明本文提出的互信息加权方法的合理性以及实用性。

4. 结论

为了解决分类问题中的属性加权问题，为不同属性赋予相应合适权重，以提高分类模型的分类准确率。本文提出了一种基于互信息的属性加权方法，并在两种经典的机器学习方法上进行了实验。将加权模型与传统模型在准确率方面的表现进行了比较。

本文的实验结果表明，我们提出的两个加权模型与传统模型相比表现相当或者更好。同时，值得一提的是，朴素贝叶斯被认为是最简单但功能最强大的分类方法之一[2]。这可能是本文的加权朴素贝叶斯分类模型与传统方法表现一样好的原因之一，需要进一步的研究来分析造成这一结果的详细原因。对于 k 近邻方法，实验结果表明，本文提出的加权 k 近邻模型的性能优于传统规则。由于 k 近邻规则是非参数分类器，因此本文提出的加权模型可以应用于任何数据集[9]。

通过实验，本文验证了基于互信息的加权方法用于机器学习模型的有效性。这种方法有以下优点：

首先，该方法以信息论为基础，权重的度量结果可信赖。

其次，该方法不会对自身稳健的分类器造成负面影响，因此具有一定普适性。

再次，该方法可以提高传统分类器的分类准确率，从而可以在实际应用中发挥重要作用。

最后，虽然该加权方法在分类准确率方面的表现良好，但其在分类效率方面的表现还有待进一步研究。

基金项目

国家自然科学基金资助项目(No. 11571009)。

参考文献

[1] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012: 122.

[2] Karabatak, M. (2015) A New Classifier for Breast Cancer Detection Based on Naïve Bayesian. *Measurement*, **72**, 32-36. https://doi.org/10.1016/j.measurement.2015.04.028

[3] Zaidi, N.A., Cerquides, J., Carman, M.J., *et al*. (2013) Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting. *Journal of Machine Learning Research*, **14**, 1947-1988.

[4] Wu, J., Pan, S., Cai, Z., *et al*. (2014) Dual Instance and Attribute Weighting for Naive Bayes Classification. *International Joint Conference on Neural Networks* (*IJCNN*), Beijing, 1675-1679. https://doi.org/10.1109/IJCNN.2014.6889572

[5] Wettschereck, D., Aha, D.W. and Mohri, T. (1977) A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithm. *Artificial Intelligence Review*, **11**, 273-314. https://doi.org/10.1023/A:1006593614256

[6] Gupta, M. (2012) Dynamic k-NN with Attribute Weighting for Automatic Web Page Classification (Dk-NNwAW). *International Journal of Computer Applications*, **58**, 34-40. https://doi.org/10.5120/9321-3554

[7] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 121-128.

[8] Mangasarian, O.L. and Wolberg, W.H. (1990) Cancer Diagnosis via Linear Programming. *SIAM News*, **23**, 1-18.

[9] Bagui, S.C., Bagui, S., Pal, K. and Pal, N.R. (2003) Breast Cancer Detection Using Rank Nearest Neighbor Classification Rules. *Pattern Recognition*, **36**, 25-34. https://doi.org/10.1016/S0031-3203(02)00044-4