样本-属性加权的朴素贝叶斯改进算法

曾文赋

(福建省福州第一中学,福建 福州 350001)

摘 要:朴素贝叶斯算法是一种简单、高效且有着广泛应用的分类方法,但在现实中,条件独立性假设影响了其分类性能。为克服该问题,给出一种改进算法——样本-属性加权的朴素贝叶斯算法。首先,对属性计算相关系数得到属性权值;其次,利用属性权结合信息熵获得样本熵权,并据此加权样本以提高泛化能力;然后,给出了样本-属性加权的朴素贝叶斯算法;最后,在 UCI 数据集上的实验结果验证了改进算法比原算法具有更好的分类性能。

关键词:朴素贝叶斯;样本-属性加权;条件独立性假设

中图分类号:TP391 **文献标识码**:A **文章编号**:1674-7720(2014)06-0062-02

Sample-attribute weighted improved naive Bayesian algorithm

Zeng Wenfu

(Fujian Province Fuzhou No.1 Middle School, Fuzhou 350001, China)

Abstract: Naive Bayesian algorithm is a simple, efficient and widely used classification method, but the conditional independence assumption affects it's classification performance in reality. The paper gives an improved algorithm——sample-attribute weighted naive Bayesian algorithm in order to overcome this problem. Firstly, the correlation coefficients of all attributes have been calculated to obtain attribute-weight. Secondly, attribute-weight and information entropy have been combined to get sample-entropy-weight, the samples have been weighted according it to enhance the generalization ability. Then, sample-attribute weighted naive Bayesian algorithm has been proposed. Finally, the experimental results on UCI data sets prove that the improved algorithm has got better classification performance than the original algorithm.

Key words: naive Bayesian; sample-attribute weighted; conditional independence assumption

分类是通过分析训练数据样本,生成分类函数或模型,通过模型将数据库中的数据映射到某一类别中,产生数据关于类别的精确描述。

朴素贝叶斯算法作为一种最简单、有效且在实际使用中很成功的分类算法,其性能可与神经网络、决策树相媲美[1]。它发源于古典数学理论,具有坚实的理论基础,与其他方法相比有较小的误差率,并广泛应用于数据挖掘、自然语言处理、医疗研究等众多领域。例如,潘志方提出一种可根据用户的网页访问记录和网上交易记录来动态地对顾客进行分类的方法[2];刘青[3]等通过不断改变 EM 算法的收敛初始条件以提高收敛效果,并结合朴素贝叶斯分类方法对未标记的中文网页进行分类;张丽伟[4]等用遗传算法对朴素贝叶斯分类算法进行改进,使之能够较好地鉴别诊断病患所属的症候,比较分析改进前后的识别效率。

朴素贝叶斯算法主要通过假设待考查的变量遵循某种概率分布,根据这些概率和已观测到的数据进行推理,作出最优决策。算法基于条件独立性假设,即假定特征向量的各分量间相对于决策变量相对独立,然而在实际应用中该假设并不现实,从而影响其分类性能。

1 朴素贝叶斯算法描述

设每个数据具有 k 个属性,用向量 $\boldsymbol{a}=[a_1,a_2,\cdots,a_k]$ 描述,其中 $a_1,a_2,\cdots,a_k$ 分别表示样本在属性 $A_1,A_2,\cdots,A_k$ 上的值。假设数据有 m 个类,分别用 $V_1,V_2,\cdots,V_m$ 来表示。给定一个样本,可得到最可能的目标值如下:

$$V_{\max}=\arg\max_j P(V_j|a_1,a_2,\cdots,a_k),V_j\in V \tag{1}$$

对一个未知数据样本 $\boldsymbol{x}=[x_1,x_2,\cdots,x_k]$,由贝叶斯定理得:

$$P(V_i|\boldsymbol{x})=\frac{P(V_i)P(\boldsymbol{x}|V_i)}{P(\boldsymbol{x})} \tag{2}$$

结合贝叶斯定理、条件独立假设和 $P(x)$ 对所有类均为常数，可判断 x 的类别如下：

$$V_{max}=\arg\max_j P(V_j|x_1,x_2,\cdots,x_k)$$
$$=\arg\max_j \frac{P(V_j)P(x_1,x_2,\cdots,x_k|V_j)}{P(\boldsymbol{x})}$$
$$=\arg\max_j \frac{P(V_j)\prod_{i=1}^{k}P(x_i|V_j)}{P(\boldsymbol{x})}$$
$$=\arg\max_j P(V_j)\prod_{i=1}^{k}P(x_i|V_j) \quad (3)$$

综上，根据朴素贝叶斯算法，对于一个未分类的样本 x，只需分别计算出 $P(V_j)$ 和 x 属于类别 V_j 的先验概率 $P(\boldsymbol{x}|V_j)$，再选出式(3)中概率最大的那个类即为 x 的类别。

2 改进策略及算法描述

由于朴素贝叶斯算法假设数据遵循某种概率分布，认为条件属性对决策属性的重要程度均相同且须满足条件独立性假设等，这些都会影响其在实际应用中的分类性能。在实际应用中，不同属性对分类影响的效果是不同的，故改进算法中考虑对不同的属性给予不同的权值，定义属性权刻画条件属性对决策属性的重要性，以克服条件独立性假设的缺陷，从而扩展朴素贝叶斯算法；同时，通过属性权结合信息熵获得样本熵权，对原始数据样本进行修正，提高算法的泛化能力。

2.1 属性权计算

训练数据集由条件属性和决策属性来描述[5]，对不同的条件属性进行加权，通过计算条件属性和决策属性间的相关系数表示两者间的相关度，得到属性权 WA。

假设 $X=(X_1,X_2,\cdots,X_k)$ 表示 k 个条件属性，Y 表示决策属性。计算 X_i 和 Y 的相关系数如下：

$$\mathrm{WA}_i=\frac{\mathrm{Cov}(X_i,Y)}{\sqrt{D(X_i)\times D(Y)}} \quad (4)$$

其中 $\mathrm{Cov}(X_i,Y)$ 为 X_i 和 Y 的协方差，$D(X_i)$、$D(Y)$ 分别为 X_i 和 Y 的方差。可知，属性权 WA_i 的值越大，表示第 i 个条件属性对分类的影响越大。

2.2 样本熵权计算

信息熵由香农所提出[6]，用来度量不确定的信息量(随机性)的大小，故计算信息熵等价于确定随机变量的分布。假设一个数据样本 $x=(x_1,x_2,\cdots,x_k)$，结合信息熵和 2.1 节中所定义的属性权计算样本熵权如下：

$$\mathrm{WS}(x)=-\sum_{i=1}^{k}x_i^{\mathrm{WA}_i}\ln x_i \quad (5)$$

通过结合属性权和信息熵定义样本熵权 $\mathrm{WS}(x)$，融合属性信息修正原始数据样本以提高泛化能力。

2.3 样本-属性加权的朴素贝叶斯算法描述

设数据集 X 中包含 n 个数据样本，每个数据样本具有 k 个属性，第 i 个样本可表示为 $X_i=(X_{i1},X_{i2},\cdots,X_{ik})$，$i=1,2,\cdots,n$。$X$ 中含有 m 个类，用 $V_1,V_2,\cdots,V_m$ 来表示。样本-属性加权的朴素贝叶斯算法步骤描述如下：

(1)对原始数据集 X 中的属性，由 2.1 节计算出属性权 WA；

(2)对原始数据集 X 中的每个样本，由 2.2 节计算出样本熵权，记为 WS；

(3)利用步骤(2)中计算获得的已融合属性信息的样本熵权 WS，对数据集 X 进行加权，得到修正后的数据集 X'，使得 X' 相比于 X 具有更好的泛化能力；

(4)对修正后的数据集 X'，使用式(6)的加权朴素贝叶斯分类模型进行分类，得到分类结果：

$$V_{max}=\arg\max_j P(V_j)\prod_{i=1}^{k}P(x_i|V_j)^{\mathrm{WA}} \quad (6)$$

其中 $P(V_j)$ 和 $P(x_i|V_j)$ 可由修正后数据集 X' 中获得，加权朴素贝叶斯分类模型的加权因子 WA_i 即为步骤(1)中计算获得的属性权。

3 实验结果与分析

实验数据采用 UCI 机器学习数据库中的 16 个数据集，在 Matlab 开发环境中完成调试，对各个数据集分别使用朴素贝叶斯算法和样本-属性加权的朴素贝叶斯算法采用十折交叉验证方式比较其分类性能。

表 1 列出了实验所使用的各个数据集名、样本数、属性数和两种算法分类的准确率。

表 1 数据集信息及两种算法比较

数据集名	样本数	属性数	朴素贝叶斯算法准确率/%	样本-属性加权的朴素贝叶斯算法准确率/%
Iris	150	4	84.67	91.33
Breast Cancer Wisconsin (Original)	699	9	97.06	97.70
Mammographic Mass	961	5	81.45	81.86
Statlog (German Credit Data)	1 000	20	69.20	69.50
Statlog (Australian Credit Approval)	690	14	66.23	77.97
Hayes-Roth	132	4	84.00	84.00
Seeds	210	7	59.52	60.95
Wine	178	13	30.04	74.84
ILPD(Indian Liver Patient Dataset)	583	10	68.18	67.53
Auto MPG	398	7	73.20	76.39
Balance Scale	625	4	68.37	68.37
BloodTransfusionServiceCenter	748	4	69.15	69.52
Pima Indians Diabetes	336	8	63.94	68.45
Ionosphere	351	34	70.83	65.00
Statlog(Vehicle Silhouettes)	846	18	47.23	61.02
Vertebral Column_2C	310	6	62.58	63.87

由上表可知，改进算法在实验中所使用的 12 个数据集分类准确率与朴素贝叶斯算法相比均有不同程度的提高；且在两个数据集上准确率相同；另外，有两个数据集的准确率低于朴素贝叶斯算法。总体上看，样本-属

性加权的朴素贝叶斯算法与朴素贝叶斯算法相比具有更好的分类性能。

本文对朴素贝叶斯算法进行改进，给出了样本-属性加权的朴素贝叶斯算法，在 UCI 数据集上进行实验，验证了改进算法相比于原算法具有更好的分类性能。

参考文献

[1] LANGLEY P,IBA W,THOMPSON K.An analysis of Bayesian classifiers[C].In:Proc of the 10th National Conference on Artificial Intelligence.MenloPark:AAA I Press,1992:223-228.

[2] 潘志方.基于朴素贝叶斯学习的电子商务网站客户兴趣分类的应用研究[J].计算机科学,2007,34(6):214-215,222.

[3] 刘青,何政.结合 EM 算法的朴素贝叶斯方法在中文网页分类上的应用[J].计算机工程与科学,2005,27(7):65-66,90.

[4] 张丽伟,段禅伦,熊志伟,等.朴素贝叶斯方法在中医证候分类识别中的应用研究[J].内蒙古大学学报,2007,38(5):568-571.

[5] 宫秀军,刘少辉,史忠植.一种增量贝叶斯分类模型[J].计算机学报,2002,25(6):645-650.

[6] Zhang Jiguo,Zhu Yongzhong.Information entropy measures for fuzziness[J].Journal of Hohai University Changzhou,2001,15(4):16-21.

(收稿日期:2013-10-26)

作者简介:

曾文赋,男,1985 年生,硕士,主要研究方向:数据挖掘与机器学习。