基于改进属性加权的朴素贝叶斯分类模型

李 方,刘琼荪
LI Fang,LIU Qiong-sun

重庆大学 数理学院,重庆 400030
College of Mathematics and Physics,Chongqing University,Chongqing 400030,China
E-mail:hszws@126.com

LI Fang,LIU Qiong-sun.Naive Bayesian classifier model based on improved weighted attributes.Computer Engineering and Applications,2010,46(4):132-133.

Abstract: To improve attribute weighted the naive Bayesian classifier model,a new measurement method of the inter-related weighted attributes is structured.The experiment proves that the naive Bayesian classifier model is superior to the classification model proposed by Zhang Shun-zhong et al.

Key words: attribute weighted;naive Bayesian;classification model;relevant measure

摘 要:构造了一种新的属性间相关性度量方法,提出了改进属性加权的朴素贝叶斯分类模型。经实验证明,提出的朴素贝叶斯分类模型明显优于张舜仲等人提出的分类模型。

关键词:属性加权;朴素贝叶斯;分类模型;相关性度量
DOI:10.3778/j.issn.1002-8331.2010.04.042 **文章编号**:1002-8331(2010)04-0132-02 **文献标识码**:A **中图分类号**:TP301

1 引言

分类是机器学习、数据挖掘方面的核心问题。近年来,从数据中提炼信息和构造可靠的分类器逐渐成为一个热门课题。分类的方法有很多,如神经网络、决策树、遗传算法、支持向量机和贝叶斯分类器等。贝叶斯分类器由于具有坚实的数学理论基础并能综合先验信息和样本数据信息,已成为分类问题的研究热点之一。而朴素贝叶斯分类器(Naïve Bayesian Classifier,NBC)是一种简单高效的分类器,在很多情况下能够取得和一些相对复杂的分类器相当的分类精度。但由于其所依赖的属性独立性假设在真实问题中往往并不成立,为此,围绕如何放松独立性假设,又能取得较好的分类效果,许多学者做了大量的工作。如 Ronald R.Yag[1]于 2004 年 5 月提出了一种通过有序加权算子作为概率乘积的权重来扩展朴素贝叶斯分类器; 程克非、张聪[2]于 2006 年 10 月提出了一种由数据导出特征加权的方法来改善朴素贝叶斯分类器;张舜仲、王树海[3]等人于 2007 年 4 月提出一种属性相关性度量公式,特征向量属于类的概率由向量相关度及其属性概率计算。

基于文献[3]提出一种改进属性加权的朴素贝叶斯分类器。实验证明,这种方法能有效提高分类效果。

2 朴素贝叶斯分类器的简介

假设数据集中有 p 个属性指标 n 个样本,第 j 个样本表示为 $X_j=(x_{j1},x_{j2},\cdots,x_{jp})$,$1\leqslant j\leqslant n$,简记 $X=(x_1,x_2,\cdots,x_p)$,x_i 表示第 i 个属性指标。假设有 m 个类别,表示为 $C_1,C_2,\cdots,C_m$,判断样本 $X=(x_1,x_2,\cdots,x_p)$ 属于类别 $C_k(1\leqslant k\leqslant m)$ 的概率可由贝叶斯公式计算。需要计算:$P(C_k/X)=\dfrac{P(X/C_k)P(C_k)}{P(X)}\propto P(X/C_k)P(C_k)$,即 $P(C_k/X)$ 的值取决于计算 $P(X/C_k)$ 和 $P(C_k)$。计算 $P(C_k)\approx s_k/n$,其中 s_k 是类别 C_k 的训练样本数,n 是总的训练样本数。

$$P(X/C_k)=\prod_{j=1}^{p}P(x_j/C_k) \tag{1}$$

判别规则:当 $P(C_k/X)>P(C_h/X)$,$k\neq h$ 时,则 $X=(x_1,x_2,\cdots,x_p)\in C_k$。

对于式(1)需要假设 p 个属性是相互独立的,但实际问题中独立性假设一般不能成立。该文基于属性相关性分析在公式(1)中加上权重系数以放松独立性的假设,即将式(1)修改为 $P(X/C_k)=w\cdot\prod_{i=1}^{p}P(x_i/C_k)$,问题的关键是恰当地构造权重系数 w。

3 属性加权贝叶斯模型

3.1 属性的加权系数

基于卡方拟合统计量的构造思想构造样本属性指标 x_k、x_j 间的相关系数。

定义 1 在类 C_i 集合中,定义样本 X 的属性指标 x_k 与 x_j 的相关系数:

$$corr_{x_k,x_j}=\frac{count(x_k,x_j)-count(x_k)count(x_j)/s_i}{count(x_k)count(x_j)/s_i} \tag{2}$$

其中,$count(x_k,x_j)$、$count(x_k)$ 分别表示在类 C_i 集合中属性对

作者简介:李方(1983-),女,主要研究方向:数理统计和模式识别;刘琼荪(1956-),女,教授,主要研究方向:智能计算,数据挖掘和应用统计。
收稿日期:2008-08-12 **修回日期**:2008-11-04

表 1　实验结果

数据集	实例数	类数	属性数	NB 正确率/(%)	CB 正确率/(%)	WB 正确率/(%)
mushroom	8 124	2	22	99.717	99.986	99.885
Tic~tao~toe	958	2	9	71.011	81.532	90.988*
Solar flare-C	1 389	2	13	80.799	84.235	84.310
Nursery	12 960	5	8	90.435	95.672	87.880
King-rook vs king-pawn	3 169	2	36	87.688	95.659	99.991*
Car evlauation	1 728	4	6	86.119	94.375	92.821
Contraceptive method	1 473	3	9	53.775	67.498	66.231
Breast cancer	699	2	9	97.039	98.299	99.451
iris	150	3	4	93.893	94.880	95.670
Pima Indians diabetes	768	2	8	77.391	84.185	85.187
平均精度				83.786 7	89.632 1	90.241 4

注：符号“*”表示分类效果提高显著。

(x_k,x_j)和 x_k 出现的频数，在 x_k、x_j 相互独立的假定下，$count(x_k)\times count(x_j)/s_i$ 可以估计理论频数 $s_iP((x_k,x_j)/C_i)$。

显然，当 x_k 与 x_j 间相互独立时，有 $P((x_k,x_j)/C_i)=P(x_k/C_i)\times P(x_j/C_i)$，根据频率近似于概率的统计思想，则 $count(x_k,x_j)s_i=count(x_k)count(x_j)$ 成立，由式(2)知，$corr_{x_k,x_j}=0$。

一般情形下，$corr_{x_k,x_j}>0$ 或<0。可以证明：$|corr_{x_k,x_j}|\leqslant 1$。因为显然 $0\leqslant count(x_k,x_j)\leqslant count(x_k)$ or $count(x_j)$，特别地，当 $count(x_k,x_j)=0$ 时，则 $corr_{x_k,x_j}=-1$；当 $count(x_k,x_j)=count(x_k)$ 时，导出 $count(x_k)=count(x_j)=\frac{s_i}{2}$，从而 $corr_{x_k,x_j}=1$。综上所述，$|corr_{x_k,x_j}|\leqslant 1$。

定义 2　在类 C_i 集合中，定义 x_k 与 x_j 的权重系数：

$$we_{x_k,x_j}=\begin{cases}|corr_{x_k,x_j}-1|, corr_{x_k,x_j}<0\\ |corr_{x_k,x_j}+1|, corr_{x_k,x_j}\geqslant 0\end{cases} \tag{3}$$

特别地，当 x_k 与 x_j 之间相互独立时，有 $we_{x_k,x_j}=1$。一般情况下，$0\leqslant we_{x_k,x_j}\leqslant 2$。

3.2　向量的加权系数

假设向量 $\boldsymbol{X}=(x_1,x_2,\cdots,x_p)$，则向量 $\boldsymbol{X}$ 的相关度可定义为：

$$we_X=\frac{P(\boldsymbol{X})}{\prod_{k=1}^{p}P(x_k)} \tag{4}$$

显然 we_X 的值越大，称向量 $\boldsymbol{X}$ 具有较大的相关性。特别地，如果 $x_1,x_2,\cdots,x_p$ 间相互独立，则 $we_X=1$。因此可由 we_X 来决定式(1)中的权重系数，但需要由样本去估计 we_X。通过分析知，向量的相关度与两两属性之间的相关度成正比，定义向量的相关度估计：

$$we_X=\begin{cases}we_{x_1,x_2}, 当\ X=(x_1,x_2)时\\ (C_p^2)^{\beta}\sqrt{\prod_{k,j=1}^{p}we_{x_k,x_j}}\ (k<j), 其他\end{cases} \tag{5}$$

其中，C_p^2 表示 we_{x_k,x_j} 相乘的个数，β 是控制参数，一般取值范围 0.1~0.3，由样本大小决定。$0\leqslant we_X<2$，选择控制参数 β 使得 we_X 的值尽可能取值于 1 的附近。

3.3　WB 分类模型

贝叶斯分类模型的关键在于求解属性 $\boldsymbol{X}$ 属于类 C_i 的概率 $P(\boldsymbol{X}/C_i)$。该文算法中，根据公式(5)计算权重系数 we_X，则 $P(\boldsymbol{X}/C_i)$ 的计算公式为：

$$P(\boldsymbol{X}/C_i)=we_X\cdot\left(\prod_{k=1}^{p}P(x_k/C_i)\right) \tag{6}$$

特别地，对于向量 $\boldsymbol{X}=(x_1,x_2,\cdots,x_p)\in C_i$，当属性 $x_1,x_2,\cdots,x_p$ 之间是独立的，则 $we_{x_k,x_j}=1$，这时取 $\beta=0$，导出 $we_X=1$，说明式(6)与式(1)是一致的。一般情况下，we_X 的值大于或小于 1。

算法步骤：

步骤 1　对于训练样本集合 D，统计类 C_i 集合中的样本数 s_i，属性 x_k 的样本数 $count(x_k)$，属性对 (x_k,x_j) 的样本数 $count(x_k,x_j)$。

步骤 2　计算先验概率 $P(C_i)=s_i/n$（n 为样本容量），计算 $corr_{x_k,x_j}$、we_{x_k,x_j} 和 $P(x_k/C_i)=count(x_k)/s_i$。

步骤 3　选取类 C_i 集合中最大的 3 个类条件概率 $P(x_k/C_i)$、$P(x_j/C_i)$、$P(x_l/C_i)$ 的 3 个属性值 x_k、x_j、x_l。扫描数据集 D，统计类 C_i 集合中包含值 (x_k,x_j,x_l) 的样本数 $count(x_k,x_j,x_l)$。类似于公式(2)和(3)，计算

$$we^{*}_{(x_k,x_j,x_l)}=\left|\frac{count(x_k,x_j,x_l)-count(x_k)count(x_j)count(x_l)/s_i^2}{count(x_k)count(x_j)count(x_l)/s_i^2}\right|$$

利用公式(5)计算 $we_{(x_k,x_j,x_l)}=(C_3^2)^{\beta}\sqrt{\prod_{k,j=1}^{3}we_{x_k,x_j}}\ (k<j)$，确定控制参数 $\beta\in(0.1\sim0.3)$，使 $\min\limits_{\beta\in(0.1\sim0.3)}\left|we_{(x_k,x_j,x_l)}-we^{*}_{(x_k,x_j,x_l)}\right|$。

步骤 4　利用公式(5)和(6)，计算 $P(\boldsymbol{X}/C_i)(1\leqslant i\leqslant m)$。

步骤 5　当 $P(C_k/\boldsymbol{X})>P(C_h/\boldsymbol{X})$，$k\neq h$ 时，则 $\boldsymbol{X}=(x_1,x_2,\cdots,x_p)\in C_k$。

4　仿真实验

为了验证前述基于加权系数的朴素贝叶斯分类器，选用了 UCI[4]机器学习数据库（ftp://ftp.ice.uci.edu/pub/machine-learning-datebases）中的 10 个数据集作为测试集，数据的基本信息如表 1。

在 PIV 1.4 GHz/256 MB 计算机上，实现了 NB 算法、CB 算法与 WB 算法，并在 UCI 数据集上完成了测试，主要采用的算法是处理离散数据，对于连续型数据实行分段处理，并转化为离散数据。由于有的数据集样本容量或属性数较多，一次测试需要较长时间，采用分割数据集的方法进行测试，其中训练集 70%，测试集 30%。其余测试集上采用十折交叉验证，并与参考文献[3]的方法进行了比较。实验结果如表 1。

仿真实验表明，WB 算法的分类正确率普遍比 NB 算法有所提高，特别是对类别数较小的情况下，分类效果较好。

5　结论及展望

文章提出的 WB 算法，在朴素贝叶斯分类器的基础上增加

（下转 141 页）