文章编号:1672 - 058X(2011)01 - 0022 - 04

基于 EM-PLS 的加权朴素贝叶斯分类算法

李雪莲
(重庆大学 数理学院,重庆 400044)

摘 要:朴素贝叶斯算法是一种简单而高效的分类算法,但是它的条件独立性假设和数据完备性要求,影响了其分类性能;在此提出了一种基于 EM 算法和偏最小二乘的加权朴素贝叶斯分类算法,实验结果验证了该算法的有效性。

关键词:加权朴素贝叶斯;EM 算法;偏最小二乘
中图分类号:TP391 **文献标志码**: A

分类问题一直是机器学习、模式识别、数据挖掘领域的核心问题。近年来,由于医疗诊断、商业智能、图像处理、故障检测等领域的发展,对分类理论提出了更高的要求。在众多的分类方法和理论中,朴素贝叶斯 NB (Naive Bayes)是目前公认的一种简单而有效的概率分类方法,由于其计算高效、精确度高,并具有坚定的理论基础而得到了广泛的应用。然而,朴素贝叶斯分类模型具有类条件独立性假设和数据完备性要求的两个先天性不足。在现实世界中,这种独立性假设和数据完备性要求经常是不满足的。因此,针对朴素贝叶斯分类的不足之处,许多学者对朴素贝叶斯分类算法进行了改进。

在数据完备性方面,陈景年等在文献[1]中通过分析著名的基于不完备数据的 RBC 分类器的不足,在 BC 方法和 EM 算法的基础上给出了一种基于不完备数据的分类器构建方法。李宏等在文献[2]中给出了一种新的基于 EM 和贝叶斯网络的丢失数据填充算法。该算法利用朴素贝叶斯估计出 EM 算法初值,然后将 EM 和贝叶斯网络结合进行迭代确定最终的更新器,同时得到填充后的完整数据集。在放宽条件独立性假设方面,Harry Zhang 等在文献[3]提出了根据属性的重要性给不同属性赋予不同的权值的加权朴素贝叶斯 WNB(Weighted Naive Bayes)分类模型,该方法更加简单可行。在此基础上,文献[4]给出了基于相关系数的属性求解算法,用相关系数度量了条件属性与决策属性之间的相关程度,却忽略了条件属性之间的关系。目前对朴素贝叶斯分类算法的改进只是单独的从数据完备性方面或者类条件独立性假设方面进行的,这种单一的改进并不能大幅提高朴素贝叶斯分类的精度。为此,提出了一种新的改进方法,采用经典的 EM 算法填补缺失数据,通过完备数据集上的朴素贝叶斯算出条件概率赋给 EM 中的期望函数作为初始参数值[2];在填补完整的数据集上,在条件属性与决策属性之间建立偏最小二乘回归方程,以回归系数作为条件属性的权值,全面提升朴素贝叶斯的分类测试能力。

1 EM 算法[5]

EM 算法是一种迭代方法,它是最常用的从不完备数据中统计参数的方法。考虑不完备数据模型,假定有两个样本空间 x 和 y,x 是完全数据,y 是观察数据。$h:x\to y$ 是从 x 到 y 的多对一映射. 以 $g(\frac{\theta}{y})$ 表示 θ 的

收稿日期:2010 - 05 - 07;**修回日期**:2010 - 06 - 28.
作者简介:李雪莲(1985 -),女,硕士研究生,从事统计模式识别以及朴素贝叶斯分类算法研究.

基于观察数据 y 的后验分布密度函数,称为观察后验分布。$f(\frac{\theta}{y,z})$ 表示添加数据 z 后得到的关于 θ 的后验分布密度函数,称为添加后验分布. $k(\frac{z}{y,\theta})$ 表示在给定 θ 和观察数据 y 下潜在数据 z 的条件分布密度函数,目的是计算观察后验分布 $g(\frac{\theta}{y})$ 的众数. 于是,EM 算法如下:首先为所求参数指派一个初值;计算不完整数据集的充分统计量的期望值;将期望的充分统计量作为真正的充分统计量参加计算。

2　PLS 回归算法[6]

偏最小二乘回归(Partial Least-squares Regression,PLS)方法是一种新型的多元统计数据分析方法,其突出优点是可以集多元回归分析、典型相关分析和主成分分析的基本功能于一体,将建模预测类型的数据分析方法与非模型式的数据认识性分析方法有机地结合起来,提供了一种多对多线性回归建模的方法,特别当变量个数很多,且都存在多重相关性时,用偏最小二乘回归建立的模型具有传统的经典回归分析等方法所没有的优点。

设因变量 $y \in \boldsymbol{R}^n$ 以及自变量集合 $X=(X_1,X_2,\cdots X_n)$,其中 $X_i \in \boldsymbol{R}^n, i=1,2,\cdots,n$。令 y 与 $X_1,X_2,\cdots X_n$ 均为标准化随机变量,即均值为零,方差为 1。按照偏最小二乘回归的第一步,首先在自变量集合 X 中提取第一个主成分 t_1,必须满足以下两个条件:t_1 和 y 应尽可能大地携带各自数据表中的变异信息;t_1 和 y 的相关程度达到最大。

$$\begin{cases} \text{maxcov}(y,t_1) = \sqrt{\text{var}(t_1)}r(t_1,y) \\ \text{s.t}\cdots t_1 = Xv_1 \cdots\cdots \| v_1 \|^2 = 1 \end{cases} \tag{1}$$

在此记 cov(·,·)是协方差算子,var(·)是方差算子,$r(\cdot,\cdot)$是相关系数算子。其中:$v_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} r^2(x_j,y)}} \begin{bmatrix} r(x_1,y) \\ \vdots \\ r(x_p,y) \end{bmatrix}$,式(1)的最优解 v_1 是对应矩阵 $X'yy'X$ 的最大特征值 θ 的标准特征向量。

在 t_1 中,关于 X_j 的线性组合系数为 $\frac{r(x_j,y)}{\sqrt{\sum_{j=1}^{p} r^2(x_j,y)}}$,显然,如果 x_j 与 y 的相关程度越强,在 t_1 成分中 x_j 的组合系数就越大! 故称 v_1 为第一主轴,t_1 为第一偏最小二乘主成分。下面实施 X 在 t_1 上的回归以及 y 在 t_1 上的回归,即 $\begin{cases} X = t_1P_1' + X_1 \\ y = r_1t_1 + y_1 \end{cases}$,其中,$P_1, r_1$ 是回归系数,记 X_1, y_1 为残差矩阵。

用 X_1, y_1 取代 X, y 重复上述步骤,提取第二个偏最小二乘主成分 t_2。依次类推,可用交叉有效性确定偏最小二乘回归中主成分 t_h 的提取个数,停止迭代。在得到主成分 $t_1,t_2,\cdots,t_m, (m<A, A=\text{秩}(X))$,得到 y 关于 t_h 的回归模型为:$y = r_1t_1 + r_2t_2 + \cdots r_mt_m + y_m$。由于 t_h 均为 X 的线性组合,则最后有:$\hat{y} = \alpha_1X_1 + \alpha_2X_2 + \cdots + \alpha_pX_p$,其中 X_j 的回归系数为 $\alpha_j = \sum_{h=1}^{m} r_hv_{hj}$。

从 α_j 的构造可以看出,如果 X_j 在构造 t_h 时贡献越大(即 v_{hj} 越大),而 t_h 在解释 y 时的作用越大(即 r_h 越大),则 X_j 在最终偏最小二乘回归模型中的回归系数就会越大。从而利用偏最小二乘原理,在条件属性 $X_1, X_2, \cdots, X_n$ 与决策属性 C 之间建立回归方程,最终可写成标准化形式:$C = \alpha_1X_1 + \alpha_2X_2 + \cdots + \alpha_nX_n$。由回归方程构建原理可知,回归系数 α_i 反映了条件属性 X_i 与决策属性 C 之间线性关系的强弱,是相关关系密切程度的重要指标。

3 分类模型与算法实现

3.1 朴素贝叶斯分类测试模型

贝叶斯分类是一种基于统计方法的分类模型,贝叶斯定理是贝叶斯学习方法的理论基础。朴素贝叶斯分类模型在贝叶斯定理的基础上,通过条件独立性假设,降低计算开销,预测未知数据样本属于最高后验概率的类。设每个数据样本用一个 n 维特征向量 $X=\{x_1,x_2,\cdots,x_n\}$ 表示,分别描述在 n 各属性 $X_1,X_2,\cdots,X_n$ 上的值。假定有 m 个类 $\{C_1,C_2,\cdots,C_m\}$,对一个未知类别的样本 X,朴素贝叶斯模型先分别计算出 X 属于每一个类别 C_i 的概率 $p(X|C_i)p(C_i)$,然后选择其中概率最大的类别作为其类别,即朴素贝叶斯分类模型为 $C_{\mathrm{NB}}(X)=\operatorname{argmax} p(c)\prod_{i=1}^{n}p(x_i\mid c)$。

3.2 加权朴素贝叶斯分类测试模型

由于在实际中难于满足朴素贝叶斯条件独立性的假设,可给不同的属性赋不同的权值使朴素贝叶斯得以扩展,则加权朴素贝叶斯模型为 $C_{\mathrm{WNB}}(X)=\operatorname{argmax} p(c)\prod_{i=1}^{n}p^{w_i}(x_i\mid c)$。其中,$w_i$ 代表属性 X_i 的权值,属性的权值越大,该属性对分类的影响就越大。加权朴素贝叶斯的关键问题就在于如何确定不同属性的权值。

3.3 EM-PWNB 算法实现

EM-PWNB 算法的基本原理是首先利用 EM 算法为缺失数据估计参数,补充完整分类集;然后在完备数据集上,建立基于偏最小二乘回归的加权朴素贝叶斯分类器。算法的具体实现如下:

(1) 运行 EM 算法,补齐分类数据集。E-step 计算不完备数据集的充分统计量的期望值。M-step 将期望统计量作为真正的充分统计量参加计算,求条件概率的极大似然估计值。

(2) 分类器的构造。概率表的学习,根据训练样本集,分别计算类 C_j 下属性 X_i 取值为 x_i 出现的概率以及类 C_j 出现的概率。权重的学习,在 X_i 与 C 之间建立偏最小二乘回归方程,对回归系数做归一化处理,将处理后的系数 w_i 作为属性 X_i 的权重。生成 PLS 的加权朴素贝叶斯概率表及属性权值表,即所需要的 PLS 的加权朴素贝叶斯分类器。

(3) 分类。调用概率表及属性权值列表,得出分类结果。

4 仿真测试实验

为了验证算法的有效性,利用 UCI 数据库[7]中的数据集对算法进行了仿真测试实验。为了比较算法的效果,在相同的硬件设施下,采用交叉法,对此算法和朴素贝叶斯分类算法进行了测试。实验首先对数据集运行 EM 算法进行缺失值填充,然后在完整的数据集上建立 PWNB 分类器,并对数据集采取十折交叉验证。每个数据集在分类器上共训练测试十次,取十次实验的平均值作为实验的测试结果。列表对比 NB 与 EM-PWNB 的正确率,实验结果见表 1。

表 1 实验数据集及分类结果 %

数据集	NB	EM-PWNB
Breast Cancer	92.17	96.23
Car Evaluation	85.56	93.52
Iris.	93.33	95.11
Nursery	90.45	94.98
German	76.73	82.29
Tic-Tac-Toe.	71.08	83.22
Letter	74.28	92.23
K-R vs K-P	87.31	95.85
平均	83.86	91.68

从表 1 可以看出基于 EM 算法的偏最小二乘加权朴素贝叶斯分类算法(EM-PWNB)相比原始的朴素贝叶斯分类算法,因一方面考虑了不完备数据的填充,相对原朴素贝叶斯的直接忽略,充分利用了数据集的原始信息,另一方面考虑了条件属性与决策属性之间的相关关系,放松了原朴素贝叶斯的基本假设,从而有效地提高了分类的正确率,从而达到更好的分类效果。

5 结论与展望

在此提出的 EM-PLS 分类算法,经实验验证具有很好的分类效果。此算法在原朴素贝叶斯分类器的基础之上,不仅考虑了不完备数据的填充,充分利用了原始数据集的信息,而且利用 PLS 增加了权重系数,放松了原朴素贝叶斯的条件独立性假设,提高了分类效果。在此只是简单的借助 EM 算法填充缺失数据,利用偏最小二乘回归,描述属性间的关联程度,这种应用显然不能充分利用它们的优良性质,如何进一步将 EM 算法和偏最小二乘应用在朴素贝叶斯分类中,以便更好的提升分类器的分类能力,将是进一步继续研究的内容。

参考文献:

[1] 陈景年,黄厚宽,田凤占,等. 一种基于不完整数据的朴素贝叶斯分类器[J]. 计算机工程,2006,32(7):86-88

[2] 李宏,阿玛尼,李平. 基于 EM 和贝叶斯网络的丢失数据填充算法[J]. 计算机工程与应用,2010,46(5):123-125

[3] ZHANG H,SHENG S. Learning weighted Naive Bayes with accurate ranking[C]. // Proceedings of the 4th IEEE International Conference on Data Mining,2004. 567-570

[4] 张卫明,王波,张斌. 基于相关系数的加权朴素贝叶斯分类算法[J]. 东北大学学报:自然科学版,2008,(29)7:952-955

[5] 张连文,郭海鹏. 贝叶斯网引论[M]. 北京:科学出版社,2006

[6] 王惠文. 偏最小二乘回归方法及应用[M]. 北京:国防工业出版社,1999

[7] BLAKE C, MERZ C. UCI repository of machine learning databases[R/OL]. University of Califomia, Irvine, Department of Information and Computer Science,1998. http://www. ics. uci. edu/_mleam/MLRepository. html

[8] FRIED N,GEIGER D,GOLDSZMIDT M,etal. Bayesian network classifiers[J]. Machine Learning,1997,29(2-3):131-163

[9] 程克非,张聪. 基于特征加权的朴素贝叶斯分类器[J]. 计算机仿真,2006,23(10):92-94

[10] 秦锋,任诗流,程泽凯. 基于属性加权的朴素贝叶斯分类算法[J]. 计算机工程与应用,2008,44(6):107-109

Weighted Naive Bayes Classification Algorithm Based on EM-Partial Least Squares

LI Xue-lian

(College of Mathematics and Physics, Chongqing University, Chongqing 400044, China)

Abstract: Naive Bayes algorithm is a simple and effective classification algorithm. However, its classification performance is affected by its conditional attribute independence assumption and request of complete data. This paper proposes a weighted Naive Bayes classification algorithm based on EM algorithm and partial least squares. Experimental results show its validity.

Key words: Weighted Naive Bayes; EM Algorithm; partial least squares

责任编辑:代晓红