第34卷 第21期 计 算 机 工 程 2008年11月
Vol.34 No.21 Computer Engineering November 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)21—0187—03 文献标识码：A 中图分类号：TP18

基于代价敏感贝叶斯网络的烟叶感官质量评价

高妍方，赵青松，陈英武
(国防科技大学信息系统与管理学院，长沙 410073)

摘 要：贝叶斯网络在判别分类中具有很多优势，应用贝叶斯网络对烟叶感官质量进行预测和评价。一些烟叶质量指标的误分类代价不同，提出一种代价敏感贝叶斯网络。通过生成准则学习代价敏感贝叶斯网络的结构，进行代价敏感参数估计。应用代价敏感贝叶斯网络对一组烟叶进行感官质量预测和评价，结果表明了代价敏感贝叶斯网络在烟叶质量感官评价中的有效性。

关键词：贝叶斯网络；代价敏感损失；感官质量评价

Tobacco Smoking Quality Evaluation Based on Cost-sensitive Bayesian Networks

GAO Yan-fang, ZHAO Qing-song, CHEN Ying-wu
(College of Information Systems and Management, National University of Defense Technology, Changsha 410073)

【Abstract】 Bayesian networks have many merits which are applied in data mine, tobacco smoking quality is predicted and evaluated using Bayesian networks. But some important smoking quality has unequal classification cost. Cost-sensitivity Bayesian network is proposed accordingly. Structure of the cost-sensitivity Bayesian networks is learned using generative criterion. Then parameters of the cost-sensitivity Bayesian networks are evaluated based on cost-sensitivity loss function. A Bayesian network is learned to form a tobacco smoking quality dataset. Experimental results show that cost-sensitivity Bayesian network is feasible to evaluate tobacco smoking quality.

【Key words】 Bayesian networks; cost-sensitive loss; smoking quality evaluation

1 概述

烟叶的感官质量和化学成分是评价烟叶质量的2个重要方面。目前为止，烟叶的感官质量评价还是依靠人工评吸。感官评吸鉴定经验性和技术性强，评估专家往往受知识结构、经验、情绪、环境等条件的影响，评估结果难免存在主观性和随意性。烟叶的感官质量是化学成分在烟气特征上的表现[1]，为了定量地进行烟叶感官质量评价，烟叶化学成分与其感官质量之间的关系一直是相关领域的重要研究课题[1-2]。这些研究大都采用传统的统计方法进行。但是烟草成分很复杂，它对人的感官刺激与人的主观感受之间的关系复杂微妙，因而无法建立化学成分与烟叶感官质量之间确定的数学模型。而传统的统计方法只能分析影响烟叶质量的相关因素，给出影响程度，却无法直接给出评估结果。为了减轻人工评价的负担，提高产品的稳定性，应用智能技术进行产品感官质量辅助评价是一个重要的发展趋势[3-4]。

贝叶斯网络是一种能够对具有统计规律的不确定性系统进行建模和推理的方法，该方法广泛地应用于判别预测问题。贝叶斯网络用于判别预测时具有传统统计方法所无法比拟的优点[5]。为此，本文应用贝叶斯网络对烟叶感官质量进行评价。

对于烟叶的某些质量指标，不同等级烟叶之间的误判代价不相等。当将高等级误判为低等级时会提高成本，而将低等级误判为高等级时则会降低产品的质量，甚至影响产品的声誉。常用的生成贝叶斯网络和判别贝叶斯网络不能够处理这种情形。为此，提出了一种代价敏感贝叶斯网络，用来进行烟叶感官质量评价问题。

2 贝叶斯网络

2.1 贝叶斯网络

贝叶斯网络[6]是表示变量间概率依赖关系的有向无环图 $B=\langle G,\Theta\rangle$，它描述了一组随机变量的联合概率分解特性。其中，$G$ 表示有向无环图，表示了变量间的依赖和独立关系；Θ 表示网络结构中的参数集合，定量地刻画了变量对其父节点的依赖关系。图中的每个节点对应一个随机变量，在这里对此不区分，均用 $X=\{X_1,X_2,\cdots,X_m\}$ 表示，而变量的取值表示为 $x=\{x_1,x_2,\cdots,x_m\}$。记节点 X_i 的取值个数为 r_i；节点 X_i 的父节点组合为 π_i；父节点组合的个数为 q_i。非根节点 X_i 所附的是条件概率分布 $P(X_i\mid\pi_i)$，而根节点 X_r 所附的是边缘概率分布 $P(X_r)$。贝叶斯网络的联合概率分布可以表示为

$$P(X_1,X_2,\cdots,X_n)=\prod_{i=1}^{m}P(X_i\mid\pi_i)$$

给定一组关于变量 $X_1,X_2,\cdots,X_m$ 的完整 i.i.d 数据集 $D=\{D_1,D_2,\cdots,D_d,\cdots,D_n\}$，其中，$D_d=(x_{d,1},x_{d,2},\cdots,x_{d,m})$。这时该数据集的对数似然函数为

$$LL(B\mid D)=\frac{1}{n}\sum_{d=1}^{n}\ln P_B(D_d)=\sum_{d=1}^{n}\sum_{i=1}^{m}\ln P_B(x_{d,i}\mid\pi_{d,i}) \quad (1)$$

当贝叶斯网络作为分类器时，假设最后一个变量为类别

基金项目：高等学校博士学科点专项科研基金资助项目(20059998019)

作者简介：高妍方(1979－)，女，博士研究生，主研方向：概率图模型，系统建模与决策；赵青松，讲师；陈英武，教授、博士生导师

收稿日期：2008-08-01 **E-mail：**gaoyanfangnudt@hotmail.com

变量 $C \in \{1,2,\cdots,|C|\}$，用 $|C|$ 表示所有类别的个数。那么，条件对数似然函数为

$$LCL(B \mid D) = \ln \prod_{d=1}^{n} P_B(C = c_d \mid X_{1:m-1} = x_{d,1:m-1}) = \sum_{d=1}^{n} \ln P_B(C = c_d \mid X_{1:m-1} = x_{d,1:m-1}) \tag{2}$$

2.2 生成贝叶斯网络

如果数据 D 是从贝叶斯网络 B 的联合概率分布 $P_B(X)$ 中抽样得到，则称 B 是 D 的生成模型，称 $P_B(X)$ 是 D 的生成分布[6]。

生成贝叶斯网络的目标是寻找联合分布 $P_B(X)$ 的最优表示模型。对数似然函数是一个模型对数据匹配程度的表示。BIC方法与BDe方法是常用的结构学习方法[6]。BIC方法基于对数条件似然进行模型选择，BDe方法则应用边际似然。最常用的生成参数估计方法包括最大似然法和贝叶斯法等。最大似然法将变量看作一个常数，而贝叶斯法将参数视为随机变量，可以融合参数的先验分布。由于贝叶斯网络的似然函数可以分解为关于贝叶斯网络局部模型的乘积，因此生成学习方法计算效率很高。

2.3 判别贝叶斯网络

生成学习的目标与判别预测问题直接将变量的条件概率作为目标相偏离，影响了贝叶斯网络分类器的性能。因此，贝叶斯网络的判别学习逐渐得到了关注。

判别贝叶斯网络结构学习方法以对数条件似然或者以分类器分类性能的评价标准作为模型选择的准则。如Grossman等提出的CMDL方法[7]，是基于对数条件似然的一种判别贝叶斯网络结构学习方法；而Pernkopf等则应用最小化分类误差准则学习贝叶斯网络结构[8]。判别贝叶斯网络参数通过最大化对数条件似然学习得到。朴素贝叶斯网络的最大化对数条件似然的参数估计等价于Logistic回归的参数估计问题。ELR算法[9]是一种能够估计标准贝叶斯网络的方法，是应用Logistic方法估计NB参数方法的扩展。

3 代价敏感贝叶斯网络

本文提出的代价敏感贝叶斯网络通过生成方法学习结构，而应用代价敏感方法估计参数。代价敏感参数估计方法基于一种代价敏感损失函数进行学习。

3.1 代价敏感损失函数

在代价敏感分类中，通常采用的分类代价矩阵(cost matrix)将正确分类的代价设为 0。本文认为相对于错误分类的情形，正确分类是有收益的。那么，样本正确分类的代价为一个负值。2 种代价矩阵分别为：

定义 1(分类代价矩阵)[10-11]：对于一个含有 $|C|$ 个类别的样本，分类代价矩阵是一个 $|C| \times |C|$ 的矩阵 $\boldsymbol{Cost}_{|C|\times|C|}$。给定类别 $i, j \in C$，$Cost[i, j] \geq 0$ 表示将第 j 类样本误分为第 i 类样本的代价，且 $Cost[i,i] = 0$。通常情况下 $Cost[i, j] \neq Cost[j, i]$。

定义 2(扩展的分类代价矩阵)：符号表示同定义 1。扩展的代价矩阵在 $i \neq j$ 时满足 $Cost[i, j] \geq 0$，且 $Cost[i,i] \leq 0$。

为了叙述方便，只讨论二分类问题。对于多分类问题是一个直接简单的扩展。设两个类别表示为 $\{1,2\}$，若 $\alpha \in \{1,2\}$ 为其中的一个类别，则另外一个类可以记为 $\beta = 3 - \alpha$。这时，定义如下的代价敏感损失函数：

定义 3(代价敏感损失函数)：给定数据集合 $D=\{x_1, x_2, \cdots, x_d, \cdots, x_n\}$，$x_d = (x_{d,1}, x_{d,2}, \cdots, x_{d,m-1}, c_d)$ 为其中的一个样本，并且 $c_d \in \{1,2\}$。对于给定的贝叶斯网络 B，定义一个样本的分类损失为

$$P_B(c_d \mid x_{d,1:m-1})^{Cost(c_d,c_d)} \cdot P_B(3-c_d \mid x_{d,1:m-1})^{Cost(3-c_d,c_d)} \tag{3}$$

那么样本集合 D 的总体分类损失的对数为

$$CSL(B, D, Cost) = \sum_{i=1}^{n} Cost(c_d, c_d) \cdot \ln(P_B(c_d \mid x_{d,1:m-1})) + Cost(3-c_d, c_d) \cdot \ln(P_B(3-c_d \mid x_{d,1:m-1})) \tag{4}$$

式(4)从形式上可以看作对数条件似然损失函数的扩展。当 $Cost(\alpha,\alpha) = Cost(\beta,\beta)$ 时，该式右边第 1 项是条件对数似然损失函数的倍数，这时第 2 项则可以看作对数条件似然损失函数的罚项。称考虑了分类代价的损失函数为代价敏感损失(Cost-Sensitive Loss, CSL)函数。

对于分类损失较大的类别，当该类别的后验概率越小时，代价敏感损失函数越大。为了最小化代价敏感损失，选择的参数会使得分类器的分类结果趋向于误分类代价较大的类。

3.2 贝叶斯网络的代价敏感参数学习

应用代价敏感损失函数进行参数估计，实际上是一个有约束优化问题：

$$\min CSL(B, D, Cost) \quad st \begin{cases} 0 < \theta_{ijk} < 1 \\ \sum_{k=1}^{|X_i|} \theta_{ijk} = 1 \end{cases} \tag{5}$$

有线性约束的非线性函数优化问题可以通过梯度投影法、简约梯度法等进行求解。

本文应用 ELR 算法中的变换方法，首先通过对数变换得到参数 $\beta_{ijk} = \ln \theta_{ijk}$，并取 $\theta_{ijk} = \frac{e^{\beta_{ijk}}}{\sum_j e^{\beta_{ijk}}}$。由于 β_{ijk} 的取值范围为实数，因此这种变换将式(5)转换为无约束优化问题。然后，通过共轭梯度法求解 β_{ijk} 的最优解。变换后的代价损失函数的梯度为

$$\nabla CSL(B, D, Cost) = \left\langle \frac{\partial CSL(B, D, Cost)}{\partial \beta_{ijk}} \right\rangle_{i,j,k} \tag{6}$$

$CSL(B, D, Cost)$ 关于 β_{ijk} 的偏导数为

$$\frac{\partial CSL(B, D, Cost)}{\partial \beta_{ijk}} = \sum_{d=1}^{n} \left[\begin{array}{l} Cost(c_d, c_d) \frac{\partial \ln(P_B(c_d \mid x_{d,1:m-1}))}{\partial \beta_{ijk}} \\ +Cost(3-c_d, c_d) \frac{\partial \ln(P_B(3-c_d \mid x_{d,1:m-1}))}{\partial \beta_{ijk}} \end{array} \right] \tag{7}$$

而 $\ln P(c_d \mid x_{d,1:m-1})$ 关于 β_{ijk} 的偏导数为[9]

$$\frac{\partial \ln P(c_d \mid x_{d,1:m-1})}{\partial \beta_{ijk}} = [P(x_{d,i}, \pi_i \mid x_{d,1:m-1}, c_d) - P(x_{d,i}, \pi_i \mid x_{d,1:m-1})] - \theta_{ijk}[P(\pi_i \mid c_d, x_{d,1:m-1}) - P(\pi_i \mid x_{d,1:m-1})] \tag{8}$$

4 算例

4.1 数据及其预处理

烟叶化学成分繁多，不可能建立所有化学成分和感官质量之间的影响关系，选择卷烟设计中评价产品质量最常用的质量指标。其中包含的常规化学指标为总糖、还原糖、总氮、烟碱、蛋白质等，感官指标选取香气质。选用 200 条烟叶质量数据，随机取其中的 160 条作为训练数据，其余的作为测试数据。用Weka系统[12]离散化连续变量，其中香气质离散为“高”和“低”两种等级。分类代价矩阵设置为[-1 ,5; 1, -1]。

4.2 实验结果

烟叶质量的代价敏感贝叶斯网络应用 BIC 准则学习结构，应用代价敏感方法估计参数，即通过 BIC+CSL 的方法构

建贝叶斯网络模型。将预测的结果与相应的生成贝叶斯网络(BIC+LL)和判别贝叶斯网络(BIC+LCL)的预测结果进行比较。10 次计算的平均值如图 1~图 3 所示。

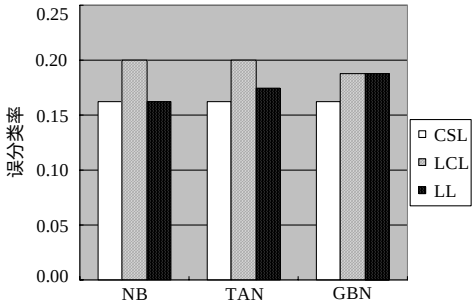

图 1　总体误分类率

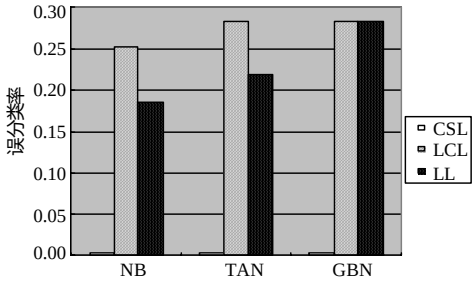

图 2　高代价类别误分类率

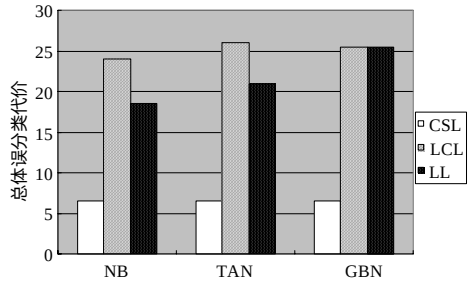

图 3　总体误分类代价

从图 1~图 3 可以看出，考虑了误分类代价的代价敏感贝叶斯网络在进行感官质量预测时，不一定具有最小的总体误分类率，但是却大大降低了较高分类代价烟叶样本的误分类率和烟叶样本的总体误分类代价。

5　结束语

本文提出了一种代价敏感贝叶斯网络用于评价烟叶的感官质量。首先通过生成准则学习代价敏感贝叶斯网络的结构；然后基于扩展的分类代价矩阵提出了一种代价敏感损失函数，并应用共轭梯度法求解参数。在一个烟叶质量数据集上的实验表明，代价敏感贝叶斯网络大大降低了总体误分类代价和高代价类别的误分类率。这表明代价敏感贝叶斯网络适合解决误分类代价不相等时的烟叶感官质量评价问题，为烟叶感官质量评价提供了一种客观的方法。

参考文献

[1] 王允白, 王宝华, 郭承芳. 影响烤烟评吸质量的主要化学成分研究[J]. 中国农业科学, 1998, 31(1): 89-91.

[2] 闫克玉, 王建民, 屈剑波, 等. 河南烤烟评吸质量与主要理化指标的相关分析[J]. 烟草科技, 2001, (1): 5-9.

[3] Munoz A M. Sensory Evaluation in Quality Control: An Overview, New Developments and Future Opportunities[J]. Food Quality and Preference, 2002, 13(6): 329-339.

[4] 郭　骏, 潘　申, 胡小建. 基于灰度形态学的烟叶图像边缘检测[J]. 计算机工程, 2007, 33(21): 163-165

[5] Chickering D. Learning Equivalence Classes of Bayesian Networks Structures[C]//Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann, 1996.

[6] 张连文, 郭海鹏. 贝叶斯网引论[M]. 北京: 科学出版社, 2006.

[7] Grossman D, Domingos P. Learning Bayesian Network Cclassifiers By Maximizing Conditional Likelihood[C]//Proceedings of ACM International Conference. Banff, Canada: ACM Press, 2004: 46-53.

[8] Pernkopf F. Bayesian Network Classifiers Versus Selective k-NN Classifier[J]. Pattern Recognition, 2005, 38(1): 1-10.

[9] Greiner R. Structural Extension to Logistic Regression-discriminative Parameter Learning of Belief Net Classifiers.pdf [J]. Machine Learning, 2005, 59(3): 297-322.

[10] Elkan C. The Foundation of Cost Sensitive Learning[C]// Proceedings of the 17th International Joint Conference on Atificial Intelligence. Seattle, Washington, USA: AAAI Press, 2001: 973-978.

[11] 凌晓峰, Sheng V S. 代价敏感分类器比较研究[J]. 计算机学报, 2007, 30(8): 1203-1212.

[12] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations[M]. San Francisco, USA: Morgan Kaufmann, 2005.

（上接第 162 页）

合法律规定的，本文所提出的电子证据安全保护策略可对非信任网络环境下的潜在电子证据进行安全保护，达到上述目标。在假设 SHA-1 和 RSA 算法安全的前提下，本文的信息记录产生方法可安全分散存储于取证服务器，通过还原和检测方法可以证明电子证据的有效性。这对电子证据的安全和有效性鉴定具有重要意义。

参考文献

[1] Sommer P. Digital Footprints: Assessing Computer Evidence[J]. Criminal Law Review, 1998, (12): 61-78.

[2] David W J, Calvert S. Digital Evidence[J]. Communications of the ACM, 2002, 45(4): 128.

[3] Sommer P. Computer Forensics: An Introduction Elsevier Advanced Technology[C]//Proc. of the 9th World Conference on Computer Security Audit and Control. London, UK: Elsevier, 1992.

[4] Schneier B, Kelsey J. Secure Audit Logs to Support Computer Forensics[J]. ACM Transactions on Information and System Security, 1999, 2(2): 159-176.

[5] 董晓梅, 王大玲, 于　戈, 等. 电子证据的获取及可靠性关键技术研究[J]. 计算机科学, 2004, 31(6): 143-145.

[6] 綦朝晖, 孙济洲, 郭琳琳. 一种支持计算机取证的信息一致性方案[J]. 计算机工程, 2006, 32(11): 172-176.

[7] Rabin M O. Efficient Dispersal of Information for Security, Load Balancing, and Fault Tolerance[J]. Journal of ACM, 1989, 36(2): 335-348.