面向多层次知识表达的贝叶斯分类模型研究

王利民 李雄飞 徐沛娟
(吉林大学计算机科学与技术学院 长春 130012)[1]
(吉林大学符号计算与知识工程教育部重点实验室 长春 130012)[2]

摘 要 提出多模式贝叶斯分类算法,由变量值之间的条件独立和条件相关性推断因果关系,根据每个完整随机样本而非整个样本空间构造子模式。结合局部计算近似推理进行概率密度和条件概率分布估计,在此基础上采用后离散化策略自动确定连续变量边界。在UCI机器学习数据集上的实验结果证明了该算法的合理性和有效性。

关键词 贝叶斯网络,多模式,后离散化策略,局部计算

Research on Bayesian Classification Model for Multi-level Representation

WANG Li-min LI Xiong-fei XU Pei-juan
(College of Computer Science and Technology, Jilin University, Changchun 130012, China)[1]
(Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China)[2]

Abstract A multi-schema Bayesian classification algorithm was proposed to solve the problem of discretization assumption and graph representation. By reasoning the conditional independence and dependence between attribute values, a submodel was constructed for each complete random instance rather that the whole instance space. The boundary of continuous attribute was decided automatically based on post-discretization strategy, the joint probability density and conditional probability were estimated based on marginal computation. The experimental study on the UCI data set shows that, this algorithm overcomes the restrictiveness of traditional TAN and can describe the marginal dependency of mixed-mode data more intuitively and accurately.

Keywords Bayesian network, Multi-schema, Post-discretization strategy, Marginal computation

1 引言

随着计算机应用和Internet的日益普及,人们利用信息技术生产和搜集数据的能力大幅度提高,无数个数据库被用于商业管理、政府办公、科学研究和工程开发等。但是现代信息管理系统的处理方式均基于数据统计技术,对指定的数据集合进行简单的数字处理,而不能对这些数据所包含的内在信息进行提取,使得隐藏在大量数据中的知识(诸如属性之间的相互约束关系等)远远没有得到充分的挖掘和利用。以指数速度增长的海量数据库信息与人们从中获取的知识之间形成强烈的反差,迫切需要改进现有的信息处理手段。

贝叶斯网络[1,2]由路径分析(因果推理链)、因果模式、影响图等逐渐演变而来。它结合图论和统计学方面的知识,作为一种有效的智能数据分析和处理的图形模式逐渐引起研究人员的广泛重视。随着人工智能的发展,尤其是机器学习、数据挖掘等领域研究成果的大量涌现,为贝叶斯网络的发展和应用提供了更为广阔的空间。文献[3]将含有机密数据的数据集垂直分割为相互独立的两部分,并提出隐私保密性协议,针对数据交叉子集构造贝叶斯网络。文献[4]结合统计独立性测试和互信息测度(MIT)描述子节点与父节点之间的相互作用。针对条件独立性假设绝大多数关系型数据集无法满足的问题,文献[5]提出关系型依赖网络(RDN)来表达和推理关系型数据库中的循环依赖,采用伪似然学习法近似估计联合分布。对于分布不均衡的高维数据,文献[6]借鉴细胞遗传学领域的方法处理:采用层次分解,由每个层次处理类别标注分布近似均衡的情况,结合维数约简对小类别标注进行上采样(Up-Sampling)。文献[7]通过评估候选结构的近似优异性,仅对最优结构进行精确计算,不仅可以改善搜索算法的运行时间,还能自然地将隐含节点加入到贝叶斯网络中。文献[8]引入模块化和面向对象思想,简化多模块贝叶斯网络(MSBN)的局部和全局推理的时空复杂度。文献[9]基于贝叶斯网络提出决策表系统的分解原理和方法,降低系统决策与推理的复杂性。

随着研究的深入,贝叶斯网络学习在理论和应用两方面的问题逐渐暴露出来。贝叶斯网络在知识推理方面有着得天独厚的优势,其主要特点之一是用图形模式表示变量间的依赖关系。学习到的知识隐藏在变量间的连接权值和图形中的有向边,网络结构主要从宏观上描述变量之间错综复杂的关系。但现实生活中特征变量之间的关系错综复杂,现有已知数据仅仅是真实模型的不完全体现。将所有可能情况集成在

到稿日期:2008-05-30 本课题得到国家自然科学基金(60275026)资助。
王利民(1974-),男,讲师,博士,主要研究方向为数据挖掘、贝叶斯网络等;李雄飞(1963-),男,教授,博士生导师,主要研究方向为数据库、智能网络系统等。

单一模式框架下进行表达，不仅使得结构过于复杂，还将导致严重的过拟合问题。单一模式贝叶斯网络不能随着实际情况的变化而动态地修正，这在一定程度上影响了用户对通过贝叶斯网络构建智能系统的信心。能否针对不同的情况描述变量间的相关性，将是贝叶斯决策在数据挖掘实用阶段必须解决的问题。因此，多模式贝叶斯网络成为自然选择，每个子模式结构简单，语义清晰，能从不同角度反映真实结构的各个侧面。

基于上述考虑，本文针对每个完整随机样本而非整个样本空间构造多模式分类子模型 MSTAN(Multi-Schema Tree Augmented Naive bayes)，以便体现变量之间在边界情况下的依赖关系。在学习过程中，采用非参数估计法和极大似然估计法，结合局部计算近似推理进行概率密度和条件概率分布估计。在此基础上采用文献[6]提出的后离散化策略对连续数据进行分析，以便在极小化信息损失的前提下自动确定离散边界。

2 MSTAN 分类模型

假设样本空间 T 由 n 维特征向量 $X=(X_1,X_2,\cdots,X_n)$ 和类变量 C 进行数据描述，分类问题的实质是找出特征向量 X 到类变量 C 的映射。贝叶斯以后验概率作为分类指示，即输出条件概率最大的类别标注作为目标值。如果特征变量均为离散的，研究人员提出条件独立性假设，以便简化变量间的依赖关系：

$$P(x_1,\cdots,x_n|c)=\prod_{i=1}^{n}P(x_i|c) \tag{1}$$

其中 $P(.)$ 表示离散的概率值，小写字母 x_i，c 分别表示特征变量 X_i 和类变量 C 的任意取值。

根据该假设生成的朴素贝叶斯(Naive bayes，NB)模型框架里，每个特征节点均以类节点作为父节点，而彼此之间无任何联系。Friedman 在 1997 年提出的该模型的扩展形式——树增广朴素贝叶斯(Tree Augmented Naive bayes，TAN)[7]，与 NB 中的类条件独立强假设不同，TAN 允许每个特征节点除了以类节点作为父节点外，最多还允许以一个其它特征节点作为其父节点。TAN 通过发现属性对之间的依赖关系来降低 NB 中任意属性之间独立的假设，它是在 NB 网络结构的基础上增加属性对之间的关联(边)来实现的。每对特征变量(X_i,X_j)之间的关联通过条件互信息进行描述：

$$I(X_i,X_j|C)=\sum_{X_i,X_j,C}P(x_i,x_j,c)\log\frac{P(x_i,x_j|c)}{P(x_i|c)P(x_j|c)} \tag{2}$$

定义 1 对于任意给定的随机样本$(x_1,x_2,\cdots,x_n)$，特征值对(x_i,x_j)之间的条件互信息为：

$$I(x_i,x_j|C)=\sum_{C}P(x_i,x_j,c)\log\frac{P(x_i,x_j|c)}{P(x_i|c)P(x_j|c)} \tag{3}$$

传统分类器的构造问题可以归结为在样本集含有的先验知识基础上，给出一个映射函数或模型，确定特征变量与类变量之间的映射关系，使得根据映射函数得到的分类规则覆盖尽可能多的训练样本，

Classifier Model：$X_1,\cdots,X_n\rightarrow C$

然后根据该分类模型判别测试样本所对应的类别标注。传统 TAN 虽然在处理不确定性的知识表达方面体现了明显的优越性，但它以整个变量空间和样本空间为基础进行分析，没有考虑到局部区域的特殊性。虽然可以从宏观上描述网络结构，但对于微观信息却难以精确表达。现实数据仅仅是真实模型的不完全体现，如果通过训练样本得到的分类规则不适用于测试样本，则极可能发生无法判别或者误判的情况。这是本文的基本观点和开展研究的出发点。考察表 1 所示的信息表，可以很容易看出，表中描述的映射关系可以表示为 $X_1,X_2\rightarrow C$。由于 u_5 中特征变量组合值$\{a_1,b_2\}$并没有在其它样本中出现，因此用现有知识无法判断类别标注。

表 1 信息表

	a_0	b_0	c_0	c_1
	a_0	b_1	c_0	c_2
	a_1	b_0	c_1	c_3
	a_1	b_1	c_0	c_4
	a_1	b_2	c_1	?

另一方面，大部分分类算法是内存驻留算法，通常假定数据量很小，算法的可伸缩性意味着对于海量数据而言是否具有有效的构造模型的能力。这一点在硬件性能提高且数据规模不断扩大的情况下显得很重要。

式(3)定义的互信息测度与 TAN 定义的条件互信息测度之间的根本区别在于，它并没有力图蕴含特征变量和类变量之间的所有情况，而仅仅考虑在特定条件下的关联性。因此结合训练数据集 T 所提供的信息，对于每个测试样本 $u=(x_1,x_2,\cdots,x_n)$都将会构造一个 MSTAN 子模型，描述 u 和类变量 C 之间的映射关系。

由于 b_2 没有出现在训练集中，因此 b_2 对于判别样本 u_5 的类别标注并没有提供任何有效信息。结合式(3)的计算可知，表 1 中描述的映射关系表示为 $x_1,x_3\rightarrow C$，相应的类别标注应该为 c_3。

学习 MSTAN 的过程如下：

输入：训练数据集 T 和测试样本$(x_1,x_2,\cdots,x_n)$

输出：MSTAN 子模型

1)对于测试样本中的连续变量值进行离散化处理。

2)计算任意特征值对(x_i,x_j)之间的条件互信息 $I(x_i,x_j|C)$。

3)建立以特征变量值$(x_1,x_2,\cdots,x_n)$为顶点、以类条件互信息 $I(x_i,x_j|C)$为权重的完全无向图。

4)生成该无向图的最大权重生成树。

5)任选一个特征变量值作为根节点，从该节点向外沿树为边定向。

6)以类变量值为类节点，添加从类节点到各个特征变量值节点的边。

根据最后按子模型所描述的网络结构，通过极大化类变量的后验概率 $P(c|x_1,x_2,\cdots,x_n)$ 来确定该样本所对应的类别标注：

$$C^*=\text{argmax}P(c|x_1\cdots,,x_n)$$

$$=\text{argmax}\frac{P(c)\prod_{i=1}^{n}P(x_i|c,\pi_i)}{P(x_1,\cdots,x_n)} \tag{4}$$

其中 π_i 为节点 x_i 所有非类父节点 Π_i 的取值。

3 参数估计

MSTAN 模型构造步骤 1 中需要对混合变量子集中的连续变量进行离散化处理。对于连续属性 X_i，根据贝叶斯定

理：

$$P(c|x_i)=\frac{P(c)p(x_i|c)}{p(x_i)} \tag{5}$$

由于数据的连续性，在一定的区间内类别标注是保持不变的。根据后离散化策略[10]，连续值 x_i 的最终边界通过信息增益确定：

$$Gain(x_i,B;S)\geqslant\frac{\log_2(M-1)}{M}+\frac{\Delta(x_i,B;S)}{M} \tag{6}$$

其中 S 为属性值的降序排列，M 为观测序列 S 中的样本数目。式(3)中的条件概率采用极大似然法进行估计：

$$\hat{P}(x_i|c)=\frac{N(x_i,c)}{N(c)} \tag{7}$$

由于满足条件$\{X_i=x_i,X_j=x_j,C=c\}$的样本数量较少，这必然降低参数估计的可靠性。本文结合局部计算近似推理进行多维条件概率分布估计，利用参数的先验分布来平衡其后验分布，以较好地解决这一问题。

$$\hat{P}(x_i,x_j|c)=\frac{N(c)}{N(c)+N'(c)}\cdot\frac{N(x_i,x_j,c)}{N(c)}+\frac{N'(c)}{N(c)+N'(c)}\cdot\theta(x_i,x_j|c) \tag{8}$$

其中，$\theta(x_i,x_j|c)$为 $P(x_i,x_j|c)$的先验估计，$N'(c)$代表了对先验的确信度。本文采用数据集中观测到的边际分布作为先验估计，并采用经验参数 $N'(c)=5$ 作为对先验的确信度。

核函数估计法是应用最广泛的非参数估计方法之一。与参数估计法相比，它对被估计的随机变量不做任何分布假设。本文利用它估计式(5)中的条件概率密度。假设当 $C=c$ 时，连续属性 X_i 的一组样本分布为$(x_{i1},x_{i2},\cdots,x_{in})$，则其条件概率密度估计值为

$$\hat{p}(X_i=x_{ij}|C=c)=\frac{1}{mh}\sum_{l=1}^{m}K(\frac{x_{ij}-x_{il}}{h}) \tag{9}$$

其中常数 h 为相应的带宽(bandwidth)，$K(.)$为高斯核函数

$$K(t)=(2\pi)^{-1/2}e^{-t^2/2}$$

对于高斯核函数，式(9)的估计是一致和渐进无偏的[11]。

在概率密度函数的核函数估计方法中，带宽 h 的选择对最终的估计结果有重要影响。若带宽选择过小，则概率密度函数估计曲线将会出现尖峰；若带宽过大，则会使估计结果过于平滑而掩盖掉某些重要的结构特征。Smyth 等[12]指出期望交叉熵(expected cross-entropy)可用来衡量概率密度函数 $p(X_i=x_{ij}|C=c)$的估计偏差：

$$CV_{CE}=-\frac{1}{m}\sum_{j=1}^{m}\log(\frac{1}{(m-1)h}\sum_{l=1,j\neq l}^{m}K(\frac{x_{ij}-x_{il}}{h}))$$

4 实验分析

目前大多数分析方法仅给出较为直观的准确率等实验结果，但很多现实问题域(如医学、金融)进行最终决策时还依赖于与类变量相关的概率趋势预测。本文采用概率代价函数(或对数损失)测度 $P_{cost}=\Sigma\log(P_i)$进行分析，其中 P_i 为正确给第 i 个样本分配类别标注的概率。该测度等价于测试数据集合的编码长度，相应较低的概率函数值对应着较优的整体概率预测性能。为了分析算法在不同情况下的性能差异性，实验在 UCI 机器学习数据库中的 12 个数据集上进行。数据集特性描述如表 2 所示。

表 2　数据集特性描述

数据集	样本集	连续变量	离散变量	类标注
BREAST	286	0	9	2
BREAST-WINS	699	0	9	2
BUPA	345	6	0	2
CLEVELAND	303	5	8	2
ECOLI	336	7	0	8
GERMAN	1000	3	21	2
GLASS	214	9	0	6
IMAGE	2310	19	0	7
IONOSPHERE	351	34	0	2
PIMA	768	8	0	2
STA-IMAGES	6435	36	0	6
SONAR	208	60	0	2

为了能够将 MSTAN 算法与 TAN 和 FTAN（Forest TAN)等算法进行客观比较，分两个阶段检验不同的离散化方法对分类性能的影响。第一阶段仅对 TAN 和 FTAN 算法建模时采用贝叶斯信息准则[4]对连续数据做离散化预处理。第二阶段，所有算法均采用贝叶斯信息准则，以避免离散边界的差异性所引起的信息损失对结构学习的影响。同时采用国际上通用的交叉验证法进行分析。将数据集 T 通过随机选取的方法分为训练集 LS 和测试集 TS，并令 $TS=DS\times10\%$，即随机选取原始数据集的 90%作为训练集生成分类模型，再根据该模型判断剩余样本集中的类别标注。经过 10 次交叉验证后，计算在所有测试集的判决准确度均值。实验结果如图 1 和图 2 所示。

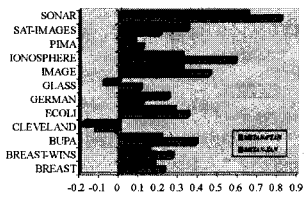

图 1　分类性能比较(离散化方法不同)

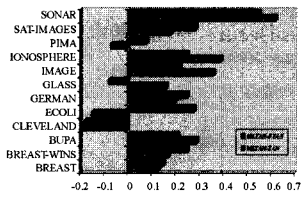

图 2　分类性能比较(离散化方法相同)

从信息论的角度来说，预离散化方法仅能利用原始数据的部分信息进行归纳学习。并且样本空间所包含的连续属性越多，预离散化引起的信息损失现象就越严重。这是 TAN 和 FTAN 算法在对数据集 SONAR 和 IONOSPHERE 的测试中判决精度较低的根本原因之一。从图 1 可以看出，MSTAN 可以将后离散化策略集成进结构学习过程中，充分利用连续属性提供的分类信息，因此在处理连续属性较多的情况时具有优势。而如果采用相同的离散化方法，在部分测试集的分类性能有所下降，如图 2 所示。

信息是对研究对象的一个统一的、全面的、不需时时改变的表达。相应地，基于信息论构造的 TAN 模型从宏观角度出发，以概率测度的权重描述属性间的相互关系。但 MSTAN 模型将互信息看作由若干个信息因子组成，每个信息因子是对研究对象在不同知识层面上的表达，数据集中的知识由分类模型的多个模式从不同角度进行表达。每个 MSTAN 模型的网络结构简单且确定。

需要指出的是，本文提出的 MSTAN 模型在分类性能和编码复杂度上的优越性是以提高计算复杂度为代价的。尤其在根据期望交叉熵估计概率密度函数时，需要的计算代价较大。

结束语 多模式分类作为传统算法的外延正在引起越来越多的研究和应用人员的重视。现实世界中混合数据的广泛存在也对数据分析的普适性提出更高的要求。本文将离散化处理和图形模式表达有机地结合在一起，构造局部性能最优的多模式贝叶斯分类模型，并给出了结构学习和参数学习的基本思路。在UCI机器学习数据集上的实验结果证明了本算法的合理性和有效性。进一步研究的内容包括先验信息的传播方式、多模式理论在回归分析、聚类和神经网络等方面的推广应用。

参 考 文 献

[1] Savakis K. Bayesian network structure learning and inference in indoor vs. utdoor image classification // Proceedings of International Conference on Pattern Recognition. 2004:479-482

[2] Peter J F. Bayesian network modelling through qualitative patterns. Artificial Intelligence, 2005, 163:233-263

[3] Zhiqiang Y, Rebecca N W. Privacy-Preserving Computation of Bayesian Networks on Vertically Partitioned Data. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(9): 1253-1264

[4] Luis M C. A Scoring Function for Learning Bayesian Networks Based on Mutual Information and Conditional Independence Tests. The Journal of Machine Learning Research, 2006, 7: 2149-2187

[5] Jennifer N, David J. Relational Dependency Networks. The Journal of Machine Learning Research, 2007, 8:653-692

[6] Boaz L, Josepha Y, Lev K. On the Classification of a Small Imbalanced Cytogenetic Image Database. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2007, 4(2): 204-215

[7] Gal E, Iftach N, Friedman N. Ideal Parent Structure Learning for Continuous Variable Bayesian Networks. The Journal of Machine Learning Research, 2007, 8:1799-1833

[8] 田凤占，张宏伟，陆玉昌，等. 多模块贝叶斯网络中推理的简化. 计算机研究与发展，2003，40(8)：1230-1237

[9] 胡小建，杨善林，胡笑旋，等. 基于贝叶斯网的决策表系统的优化分解. 计算机研究与发展，2007，44(4)：667-673

[10] Wang L M, Yuan S M. Induction of hybrid decision tree based on post-discretization strategy. Progress in Natural Science, 2004, 14(6): 541-545

[11] Silverman B W. Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability, 1986

[12] Smyth P, Gray A, Fayyad U. Retrofitting decision tree classifiers using kernel density estimation // Proceedings of the 12th International Conference on Machine Learning. Morgan Kaufmann Publishers, 1995:506-514

(上接第99页)

Realsecure，Snort基本漏报率和标准特别漏报率，如图3所示。

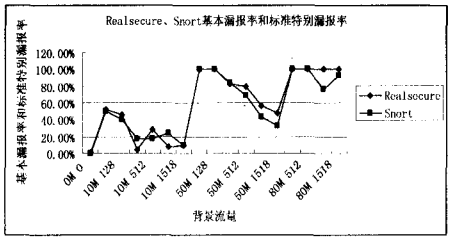

图3 Realsecure，Snort基本漏报率和标准特别漏报率

从Realsecure和Snort的两个漏报率指标可以看出：

(1)在小流量大字节的情况下，Snort的漏报率高于Realsecure。

(2)在大流量大字节的情况下，Realsecure的漏报率高于Snort。

(3)在小字节的情况下，二者漏报率都很高，甚至达到100%。但Snort的表现要比Realsecure好。

Realsecure，Snort异常特别漏报率如图4所示。

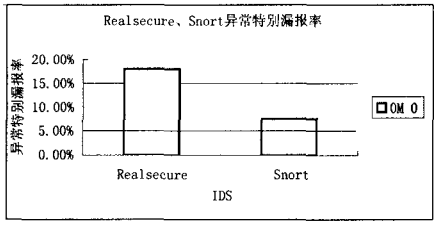

图4 Realscure，Snort异常特别漏报率

由上图可知，Realsecure比Snort的异常特别漏报率要高很多。

以上，根据基于变化流量互补测试集的入侵检测系统测试的思想，应用与之对应的指标体系，通过对Realsecure和Snort两类具有代表性的入侵检测系统的实际测试，得到了关于两者检测入侵行为的测试数据，并进行了相应的对比分析，通过比较，两者在不同的条件下表现出来的性能还是有所区别，但是总体来说各有特长，各有其发挥出色之处。应该说，符合两个系统在实际应用中展现出来的特征，以此验证了该测试方案的正确性和可行性。同时，也通过该测试结果，向使用者表明，使用单一种IDS无法全面解决检测问题，必须相互配合才能真正解决异常检测问题。

参 考 文 献

[1] 郑飞，方敏. 入侵检测技术研究 [J]. 计算机仿真，2004，21(8)：70-73

[2] 张涛，董占球. 网络攻击行为分类技术的研究 [J]. 计算机应用，2004，24 (4)：115-118

[3] 汪洋，龚俭. 入侵检测系统评估方法综述 [J]. 计算机工程与应用，2003，39(32)：171-173

[4] 张雪芹，顾春华，林家骏. 入侵检测技术的挑战与发展 [J]. 计算机工程与设计，2004，25(7)：1096-1099

[5] 蔡忠闽，孙国基，卫军胡，等. 入侵检测系统评估环境的设计与实现 [J]. 系统仿真学报，2002，14(3)：377-380