一种基于树增强朴素贝叶斯的分类器学习方法

陈 曦 张 坤*

(长沙理工大学计算机与通信工程学院 长沙 410114)

(长沙理工大学综合交通运输大数据智能处理湖南省重点实验室 长沙 410114)

摘 要：树增强朴素贝叶斯(TAN)结构强制每个属性结点必须拥有类别父结点和一个属性父结点，也没有考虑到各个属性与类别之间的相关性差异，导致分类准确率较差。为了改进TAN的分类准确率，该文首先扩展TAN结构，允许属性结点没有父结点或只有一个属性父结点；提出一种利用可分解的评分函数构建树形贝叶斯分类模型的学习方法，采用低阶条件独立性(CI)测试初步剔除无效属性，再结合改进的贝叶斯信息标准(BIC)评分函数利用贪婪搜索获得每个属性结点的父结点，从而建立分类模型。对比朴素贝叶斯(NB)和TAN，构建的分类器在多个分类指标上表现更好，说明该方法具有一定的优越性。

关键词：贝叶斯分类器；树增强朴素贝叶斯；评分函数

中图分类号：TP311.1 **文献标识码**：A **文章编号**：1009-5896(2019)08-2001-08

DOI: 10.11999/JEIT180886

A Classifier Learning Method Based on Tree-Augmented Naïve Bayes

CHEN Xi ZHANG Kun

(*School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha* 410114, *China*)

(*Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha University of Science and Technology, Changsha* 410114, *China*)

Abstract: The structure of Tree-Augmented Naïve Bayes (TAN) forces each attribute node to have a class node and a attribute node as parent, which results in poor classification accuracy without considering correlation between each attribute node and the class node. In order to improve the classification accuracy of TAN, firstly, the TAN structure is proposed that allows each attribute node to have no parent or only one attribute node as parent. Then, a learning method of building the tree-like Bayesian classifier using a decomposable scoring function is proposed. Finally, the low-order Conditional Independency (CI) test is applied to eliminating the useless attribute, and then based on improved Bayesian Information Criterion (BIC) function, the classification model with acquired the parent node of each attribute node is established using the greedy algorithm. Through comprehensive experiments, the proposed classifier outperforms Naïve Bayes (NB) and TAN on multiple classification, and the results prove that this learning method has certain advantages.

Key words: Bayesian classifier; Tree-Augmented Naïve Bayes (TAN); Scoring function

1 引言

贝叶斯网络[1]表达一种因果关系，用图模型理论和统计学知识表示属性之间的概率，在贝叶斯网络中，分类是根据类别的先验分布计算后验概率，从而选择最可能的类。贝叶斯分类器在时间和空间复杂性上具有优秀的表达能力[2]。朴素贝叶斯(Naïve Bayes, NB)分类器[3]是一种简单有效的贝叶斯网络，但由于其属性变量之间存在条件独立性假设，分类精度不佳。Friedman等人[4]提出树增强的朴素贝叶斯(Tree-Augmented Naïve Bayes, TAN)，它允许属性结点最多只能依赖于一个非类结点，综合性能良好，是学习效率与分类精度之间的一种折中。

目前关于TAN分类器的研究通常从构建合适的贝叶斯网络着手，文献[5]提出一种不确定条件互信息度量方法来学习树形贝叶斯分类网络结构；文献[6]根据条件对数似然性提出一种平均树增强朴素

收稿日期：2018-09-18；改回日期：2019-03-27；网络出版：2019-04-20

*通信作者： 张坤 zonkis2016@outlook.com

基金项目：国家自然科学基金(61772087)

Foundation Item: The National Natural Science Foundation of China (61772087)

贝叶斯；文献[7]对TAN分类器结构空间和TAN分类器结构等价类空间进行了研究，提出一个不考虑边重定向的TAN分类器学习算法。这类低阶或受限(如k阶依赖贝叶斯(k-BAN)[8,9])的贝叶斯分类模型既避免了由高维计算导致的不稳定性[10,11]，同时也增强了网络结构中属性之间的因果关系。然而，TAN模型虽然简洁高效，但在构建网络结构时并没有进行相关属性选择或引入新属性，这对TAN分类模型的分类精度有所影响。

本文在保证TAN精简结构的基础上，提出扩展的TAN分类器，额外允许TAN模型中部分属性没有父结点。考虑到属性对类贡献程度差异，采用互信息测试进行属性选择，用于确定后续每个属性结点的候选连接。并给出了利用可分解的评分函数来构建TAN模型的详细过程(此过程中的模型简记为STAN)，提出一种利用改进的贝叶斯信息标准(Bayesian Information Criterion, BIC)评分函数来构建树形贝叶斯网络分类模型(Extended Tree Augmented Naïve Bayes with the Score function, SETAN)的学习方法。通过与其它同类分类器进行对比实验，本文提出的SETAN分类模型取得了更好的分类精度。

2 基于BIC评分函数的SETAN分类器

2.1 TAN模型

2.1.1 TAN模型

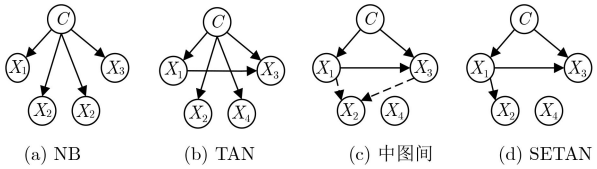
TAN分类器是一种满足Markov条件[12]的树形结构的贝叶斯网络分类器，其结构如图1(b)所示，对比图1(a)中的NB结构，它允许每个属性结点除了类父结点外，至多只有1个属性父结点。对于概率分布$P(X_1, X_2, \cdots, X_n, C)$, TAN分类器可表示为

$$\arg\max_{C(x_1,x_2,\cdots,x_n)}\left(P(C)\prod_{i=1}^{n}P\left(X_i|\Pi_i, C, G_{\mathrm{T}}\right)\right) \tag{1}$$

其中，G_{T}表示在给定类结点C的约束条件下$X_1, X_2, \cdots, X_n$的最大权重跨度树，Π_i是在最大权重跨度树中X_i的属性父结点，Π_i取值为0或1。因此，学习TAN结构首先要建立一个无向图结构，再找到合适的算法来解决最大权重生成树问题。

2.1.2 STAN(Tree-Augmented Naïve Bayes with Score function)模型

评分与搜索方法是常见的一种贝叶斯网络结构学习方法，它将结构学习转化成最优化问题，学习目标即搜索评分较高的网络结构。评分搜索的结构学习方法分为两步：网络结构评分函数和网络结构学习算法的确定。一旦定义好评分函数，贝叶斯网络结构学习问题就是一个最优化搜索问题。



(a) NB (b) TAN (c) 中图间 (d) SETAN

图 1 结构示意图

(1) 评分函数

假设给定完整训练集D，$D=\{X_1, X_2, \cdots, X_n\}$，$G$是以$X_1, X_2, \cdots, X_n$为结点的贝叶斯网络。假设数据集满足独立同分布假设，则G相对于数据集D的优劣可以用评分函数来度量。探索最佳贝叶斯网络结构，就是找到可使得评分函数最大化的一个有向无环图G。即

$$G_{\max}=\arg\max_{G\in G_x}\mathrm{Score}_D(G) \tag{2}$$

若选定的评分函数$\mathrm{Score}_D(G)$满足似然等价性和可分解性，图G_1, $G_2(G_1\neq G_2)$是在属性结点集上的两个任意图，变量之间的条件独立性相同，当且仅当$\mathrm{Score}_D(G_1)=\mathrm{Score}_D(G_2)$时，此时称$\mathrm{Score}_D$是似然等价的，可分解性则意味着有向无环图$G$可用它的局部结构表示，即

$$\mathrm{Score}_D(G)=\sum_{i=0}^{n}\mathrm{Score}_D(X_i, \Pi_i) \tag{3}$$

评分函数是影响贝叶斯网络结构学习精确度的一个主要因素，选择合理的评分函数是贝叶斯网络结构学习的一个核心问题。文献[13]提出BIC评分函数，具备可分解性和似然等价性。BIC评分函数是在样本满足独立同分布假设的前提下，用对数似然度度量结构与数据的拟合程度。具体形式为

$$\begin{aligned}\mathrm{BIC}(G_{\mathrm{T}}|D)=&\sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}m_{ijk}\lg\theta_{ijk}\\&-\frac{1}{2}\sum_{i=1}^{n}q_i(r_i-1)\lg m\end{aligned} \tag{4}$$

其中，q_i是变量X_i父结点取值组合的数目，r_i是变量X_i的取值数目，m是样本数，m_{ijk}表示X_i的父结点取j值，X_i取k值时的样本个数。$m_{ij}=\sum_{k=1}^{r_i}m_{ijk}$，$\theta_{ijk}=m_{ijk}/m_{ij}$。

(2) 模型描述

下面给出TAN分类模型结合评分搜索方法STAN的一般性表达式。给定评分函数时，在TAN无向图中，有

$$\begin{aligned}&\mathrm{Score}_D(X_i,\{C\})+\mathrm{Score}_D(X_i,\{C,X_j\})\\&=\mathrm{Score}_D(X_j,\{C\})+\mathrm{Score}_D(X_j,\{C,X_i\})\end{aligned} \tag{5}$$

由对称性可定义图中每条边(X_i, X_j)的权重

$$w(X_i,X_j)=\mathrm{Score}_D(X_i,\{C\})-\mathrm{Score}_D(X_i,\{C,X_j\}) \quad (6)$$

假设只有X_1不具备属性父结点，则可求得TAN分类器的网络结构表达式

$$\begin{aligned}G_{\mathrm{TAN}}&=\mathop{\mathrm{argmax}}_{G\in G_{\mathrm{TAN}}}\mathrm{Score}_D(G)\\&=\max_{G\in G_{\mathrm{TAN}}}\left(\sum_{i=2}^{n}\mathrm{Score}_D(X_i,\{C,\Pi_i\})+\mathrm{Score}_D(X_1,\{C\})\right)\\&=\mathrm{Score}_D(X_1,\{C\})-\min_{G\in G_{\mathrm{TAN}}}\left(-\sum_{i=2}^{n}\mathrm{Score}_D(X_i,\{C,\Pi_i\})\right)\end{aligned}$$

结合式(5)和式(6)，则有

$$G_{\mathrm{TAN}}=\sum_{i=1}^{n}\mathrm{Score}_D(X_1,\{C\})-\min\sum_{i=2}^{n}w(X_i,\Pi_i) \quad (7)$$

其中，最小式即最小生成树问题，将其最小化即可求得TAN分类网络结构，如图1(b)所示。当权重$w(X_i,X_j)\geq 0$时，移除边(X_i,X_j)。

2.2 SETAN模型

2.2.1 理论分析

由TAN的定义和式(7)可知，TAN结构限制每个属性结点X_i对其父结点有如下两种连接选择：(1)只有类父结点C；(2)具有类父结点C和一个属性父结点X_j。TAN的学习是在完全图中搜索弧空间，通过这种限制，减小了搜索空间；同时父结点的数量受限使得条件概率计算相应地减少。

然而，Greiner等人[14]通过实验证明，与数据集实际分布近似或比数据集实际分布简单的网络结构都具有一定的局限性。文献[7]已给出证明，限制父结点的数量可以有效避免具有指数复杂度的高维计算。而且即使NB网络或TAN网络相对简单，也可能由于存在冗余的结点和弧边而使得网络结构复杂化。显然，TAN结构并不能充分地表示属性结点之间的依赖关系，而且在构建网络结构时也未能去除冗余的属性结点，TAN是在维持原始属性变量集合的基础上建立低阶树形分类模型，而不是通过引入新的属性变量来放松条件独立性假设。因此，在建立TAN结构前有必要进行属性选择，既能保持模型的简洁性，同时进一步压缩结构学习过程中的搜索空间。另外，TAN结构仅仅强化了属性之间的因果关系，而没有考虑不同属性对类的贡献，这同样也降低了TAN模型的分类准确性。文献[15,16]的一系列对比实验证实了这一结论。

基于上述分析，本文进一步扩展了TAN网络结构，该网络结构相对于TAN能够更充分地表示在类约束下属性之间的依赖关系，并尝试剔除对分类模型没有贡献的属性结点。

2.2.2 SETAN模型

(1) 条件独立性(CI)测试

由香农的信息论可知，互信息可用作两个变量之间相关性度量，两个随机变量X_i和X_j之间的互信息为

$$\begin{aligned}I(X_i,X_j)&=H(X_i)-H(X_i|X_j)\\&=\sum_{x_i,x_j}P(X_i,X_j)\lg\left[\frac{P(X_i,X_j)}{P(X_i)P(X_j)}\right]\end{aligned} \quad (8)$$

当$I(X_i,X_j)=0$或小于某阈值ε，认为随机变量X_i和X_j相互独立。互信息测试又可以称为0阶CI测试。利用互信息在量化邻居结点之间的影响，比使用更多三角形结构信息的方法具有更好的性能[17]。

(2) 改进BIC评分函数

在式(4)中，第1项是模型的对数似然度，评估网络结构与数据的拟合程度。第2项是惩罚项，避免模型过拟合。θ_{ijk}表示贝叶斯网络中变量X_i的似然条件概率，且存在$0\leq\theta_{ijk}\leq 1,\sum_k\theta_{ijk}=1$。当父结点取第$j(1\leq j\leq q_i)$个值时，$\max(\theta_{ijk})$的值越接近1，表明在第$j$个取值时，父结点和子结点的因果关系越强；反之，$\max(\theta_{ijk})$越接近$1/r_i$，表明其因果关系越弱，甚至不具有因果关系。

选择相关贝叶斯模型时，结合BIC评分函数确实有可能找到具有最大后验概率的目标模型，但当目标函数无边界时会失效[18]。由式(4)中固定惩罚结构可知，对于同一个贝叶斯网络，样本数越大，网络结构的得分相应越大，分值与样本数呈单调递增关系，容易产生过拟合，影响模型性能。因此，一方面既希望类似式(7)和式(10)的评分函数最大化，但又需尽可能避免模型过拟合现象，可以添加惩罚系数ξ，重新改写BIC评分函数的惩罚项为$\xi\sum_{i=1}^{n}q_i(r_i-1)\lg m$，该惩罚项可进一步简化成$\xi\sum_{i=1}^{n}\lg m\cdot q_i\cdot r_i$。根据BIC评分函数的可分解性，可得到改进后BIC的家族评分函数为

$$\begin{aligned}\mathrm{Score}_D(G_{\mathrm{T}})&=\mathrm{BIC}((X_i,\Pi_i)|D)\\&=\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}mi_{jk}\lg\frac{m_{ijk}}{m_{ij}}-\xi\lg mq_i\cdot r_i\end{aligned} \quad (9)$$

(3) 改进TAN模型

为避免贝叶斯网络分类器的高维计算，同时为了在构建SETAN结构时去除冗余结点并减少候选父结点集的搜索空间，增强分类模型的可靠性与健

壮性，在TAN结构基础上，允许属性结点没有父结点。即具有如下额外两种选择：只有一个属性父结点；没有父结点。其中没有父结点的属性结点被视为对分类模型没有贡献的冗余结点。考虑到SETAN结构中各个属性结点X_i和类结点C的相关性不同，先对类结点和属性结点进行互信息测试，如图1(c)中间图和图1(d)SETAN结果图所示。

对于互信息大于阈值ε的结点X_i和X_j，计算并比较Score$(X_i,\{C\})$，Score$(X_i,\{X_j,C\})$，Score$(X_i,\{X_j\})$三者评分大小，添加有向弧，如结点X_3，Score$(X_3,\{X_1,C\})$最大，则有$C\to X_3$和$X_1\to X_3$；

对于互信息小于阈值ε的结点X_k，计算并比较Score$(X_k,\{X_i\})$和Score$(X_k,\{\varnothing\})$，添加有向弧，如结点X_2，不能将类结点作为父结点，且Score$(X_2,\{X_1\})$最大，则有$X_1\to X_2$；同理，对于结点X_4，则被视为对类没有贡献的冗余结点。

基于上述改动，每个属性结点不必将类结点纳入候选父结点集，则式(5)中的对称性无法成立，从而无法在式(7)中利用最小生成树算法求得SETAN结构。此时对于经过0阶CI测试后的无环图，采用BIC评分函数贪婪查找下一个局部无环图G，从而得到图1(d)所示的最终有向无环图。则有

$$G_{\text{SETAN}}=\arg\max_{G}\left(\sum_{i=1}^{k_1}\text{Score}_D(X_i,\Pi_i)+\sum_{j=1}^{k_2}\text{Score}_D(X_i,\Pi_j)\right)\tag{10}$$

其中，Π_i和Π_j分别是符合不同互信息测试的属性结点的父结点集，k_1是符合互信息大于阈值ε的属性结点个数，k_2是符合互信息小于阈值ε且对网络结构有贡献的属性结点个数，所以$k_1+k_2\le n$。因此，对于概率分布$P(X_1,X_2,\cdots,X_n,C)$, SETAN分类器的表示形式为

$$\arg\max_{C(x1,..xn)}\left(P(C)\prod_{i=1}^{n}P(X_i|\Pi_i,C,G_{\text{SETAN}})\right)\tag{11}$$

式(11)和式(1)类似，区别在于Π_i是否为SETAN结构中符合互信息测试属性X_i的父结点集，且有$\Pi_i\supseteq\{C\}$。

2.2.3 算法描述

基于BIC评分函数的SETAN分类器学习方法主要有如下改进：(1)提出SETAN网络结构，在TAN结构基础上中放松了每个属性结点的父结点选择条件，允许部分属性没有类父结点，在同等计算复杂度下提高了分类模型的可靠性；(2)结合上述属性依赖条件，采用低阶CI测试，获得各个属性的候选父结点集合，同时也获得候选的冗余结点集，压缩了评分搜索学习方法的搜索空间；(3)将类结点纳入整个网络结构学习过程中，利用改进的BIC评分函数对局部最优无环图进行贪婪查找，从而获得最终的SETAN网络结构。进一步去除无效结点，提高算法的分类精度，如表1所示。

表 1 算法描述

输入：变量集V，样本数据集D	
输出：SETAN结构	
步骤1	For each $X_i\in V$, $I(C,X_i)=$ Calc_MI(C,X_i)# 计算属性与类别之间的互信息值
步骤2	将每个互信息值$I(C,X_i)$存入数组，降序
步骤3	For each $I(C,X_i)>\varepsilon$ in List
	$S_1=S_1\cup\{X_i\}$
	Add path $C-X_i$to graph E# 若无连接边，则添加类别C到属性X_i的连接边
	$S_2=S_2\cup\{X_j\}$，$X_j\in\{I(C,X_j)<\varepsilon\}$
	Add path $X_i-X_j$to graph E# 互信息值小于阈值ε的结点则被添加到集合S_2
	Remove $I(C,X_i)$from List
	End for
步骤4	For each $E'\in E$
	Score$(E')=$ Calc_BIC(E')# 计算改进的BIC评分
	K2-Search of the optimal SETAN Structure # 利用评分函数搜索最优结构
	End for
步骤5	Return $G=(V',E')$with best BIC score

2.2.4 时间复杂度分析

从整体上看，本文提出的SETAN分类器学习方法主要分为两个部分：首先是类结点与各个属性结点之间的互信息测试。主要耗时是互信息测试$I(C;X_i)$，复杂度为$O(Nn)$，N是训练集实例数量，n是属性结点数量；第2部分是构建SETAN网络结构，主要是需要比较每个结点和其候选父结点集的连接得分，以此确定其父结点。时间复杂度是$O(Nk_1^2+Nk_1\cdot k_2)$，因为$k_1+k_2\le n$, ε一般取0.01～0.05，大多数属性结点可符合互信息测试，即$k_2\ll k_1$。因此，SETAN分类器最终可在$O(Nn^2)$内完成，和TAN分类模型的时间复杂度相同。

3 实验结果与分析

3.1 BIC评分函数惩罚系数估计

本节实验的主要目的是确定式(9)中改进后BIC评分函数的合适的惩罚系数ξ。分别采用http://www.norsys.com提供的Asia网和Alarm网进

行仿真实验。Asia网包含8个变量和8条边，Alarm网包含33个结点，46条边，样本数量均为5000。利用常见的K2算法和改进后的BIC评分函数学习贝叶斯网络结构，惩罚系数ξ分别取0.01, 0.001, 0.0001。实验结果如表2所示，A为增加边，D为确实边，R为正确边。从表2的实验结果可以看出，$\xi=0.01$时，Asia网络和Alarm网络结构缺式边数量相对比较多，没有增加边，说明惩罚系数偏大，导致数据和网络结构欠拟合；$\xi=0.0001$时，网络结构增加边相对较多，导致数据和网络结构过拟合；而当$\xi=0.001$时，各项数据比较合理，说明数据和网络结构拟合较好。

表 2　网络结构学习实验结果

ξ	Asia网			Alarm网		
	A	D	R	A	D	R
0.01	0	3	5	0	20	25
0.001	1	1	7	3	2	45
0.0001	3	0	8	15	3	45

3.2　SETAN模型分类性能评估

3.2.1　实验环境

实验数据选取UCI资源库中6个具有代表性的离散数据集、KDD Cup2008“乳腺癌早期检测问题”数据集和1987年美国众议院选举投票的数据集，数据信息如表3所示。实验环境在Windows7操作系统上进行，集成开发环境Intellij Idea, Weka 3.8，硬件配置为Intel®Core(TM)i5-2410MCPU@2.30 GHz，内存1 GB。

3.2.2　阈值ε的准确率对比

由式(8)可知，两两结点间互信息值小于阈值时，则被认为相互独立。为避免这类经验性参数的干扰[19]，分别在表3中4个数据集上进行阈值对比实验。实验采用十折交叉有效性验证的方法，对于缺失值将其作为一个单独的值来处理，实验结果取平均值，其中ε分别取0.01, 0.05和0.1。实验结果如表4所示。

图2展示了各个数据集上SETAN在不同阈值ε情况下的分类准确率。阈值ε的设定会影响两两结点是否相互独立的概率，在同一数据集上，阈值ε越小，式(10)中与类结点相关的属性结点数会相应增加；反之，阈值ε越大，可能存在的冗余结点数增加。因此，阈值ε越小，式(10)代表的网络结构得分越高，从而构建的分类效果越好。但同时，阈值ε越小，相关的结点数增加，随之计算量也会相应增加。由图可知，阈值ε取0.01和0.05时的分类准确率非常接近，因此ε取0.05较为合理。

表 3　实验数据集信息

数据集	样本数量	类别分布	属性数量	分类数量	缺失值
Balance	625	49/288/288	4	3	否
Car	1728	1210/384/69/65	6	4	否
Connect	67558	44473/16635/6449	42	3	否
Mushroom	8124	4208/3916	22	2	是
Nursery	12960	4320/2/328/4266/4044	8	5	否
SPECT	80	40/40	22	2	否
Cancer	286	85/201	9	2	否
Votes	435	168/267	16	2	是

表 4　阈值ε信息

数据集	阈值ε	平均准确率
Balance	0.01/0.05/0.10	0.915/0.914/0.910
Connect	0.01/0.05/0.10	0.767/0.764/0.760
SPECT	0.01/0.05/0.10	0.740/0.738/0.733
Cancer	0.01/0.05/0.10	0.710/0.710/0.698

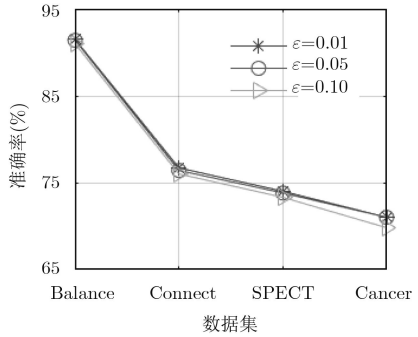


图 2 不同阈值ε的分类准确率

3.2.3　实验结果及分析

本文采用准确率(Accuracy)、召回率(Recall)、精确率(Precision)、F1值(F1-measure)、受试者工作特征曲线(Receiver Operating Characteristic, ROC)与坐标轴围成的面积(Area Under

ROC Curve, AUC)5个常见的分类指标进行性能评估，ROC曲线的横轴是假正例率(FPR)，纵轴是真正例率(TPR)。相较于Precision和Recall等衡量指标更加合理，AUC值越大说明模型性能越好。FPR和TPR二者定义如式(12)

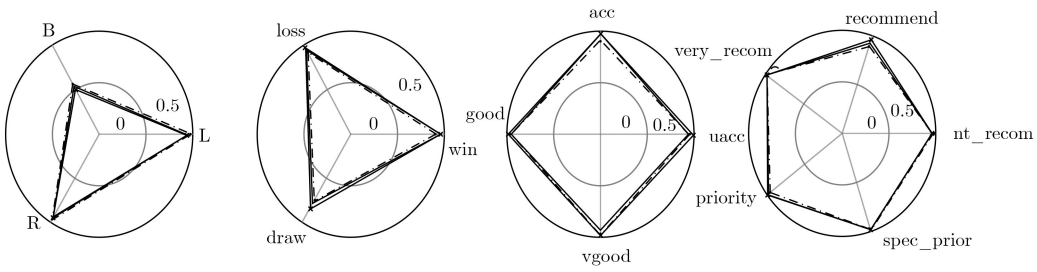
$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} \tag{12}$$

所用的低阶CI测试阈值ε取值为0.05，BIC评分函数的惩罚系数ξ取0.001。在实验中采用十折交叉有效性验证的方法，对于数据集中的缺失值，将其作为一个单独的值来处理，实验结果取平均值。表5、图3—图7分别给出了本文提出的SETAN算法与NB, TAN算法的详细实验结果。

从表5可以看出，SETAN在多分类或二分类数据集上相对有更好的分类效果；对于类别分布不均衡的数据集(如Balance, Car, Nursery, Cancer)，SETAN的各项分类指标整体上优于NB和TAN分类器；其次，SETAN分类模型也适用于不同数据

表 5 NB, TAN和SETAN的各项分类指标对比情况

数据集	算法	准确率	F1值	召回率	精确率	AUC面积
Balance	NB	0.914	0.876	0.914	0.842	0.961
	TAN	0.861	0.834	0.861	0.836	0.904
	SETAN	0.914	0.876	0.914	0.842	0.962
Car	NB	0.857	0.849	0.857	0.854	0.976
	TAN	0.908	0.911	0.908	0.92	0.983
	SETAN	0.946	0.947	0.946	0.947	0.988
Connect	NB	0.721	0.681	0.721	0.681	0.807
	TAN	0.763	0.722	0.763	0.731	0.864
	SETAN	0.764	0.724	0.764	0.735	0.866
Mushroom	NB	0.958	0.958	0.958	0.96	0.998
	TAN	0.999	1.000	0.999	1.000	1.000
	SETAN	1.000	1.000	1.000	1.000	1.000
Nursery	NB	0.903	0.894	0.903	0.906	0.982
	TAN	0.928	0.92	0.928	0.929	0.991
	SETAN	0.937	0.927	0.937	0.937	0.993
SPECT	NB	0.738	0.735	0.738	0.745	0.802
	TAN	0.713	0.709	0.713	0.724	0.668
	SETAN	0.738	0.736	0.738	0.741	0.755
Cancer	NB	0.734	0.727	0.734	0.723	0.702
	TAN	0.706	0.692	0.706	0.687	0.667
	SETAN	0.710	0.700	0.710	0.695	0.624
Votes	NB	0.901	0.902	0.901	0.905	0.973
	TAN	0.940	0.940	0.940	0.941	0.986
	SETAN	0.949	0.950	0.949	0.950	0.985



(a) Balance数据集 (b) Connect数据集 (c) Car数据集 (d) Nursery数据集

图 3 多分类数据集的AUC polar图对比

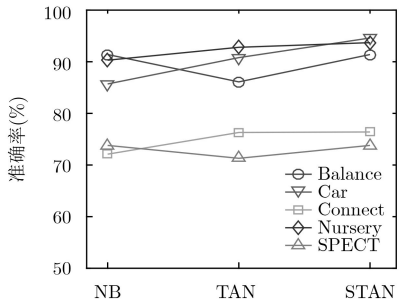


图 4 平均分类准确率对比图

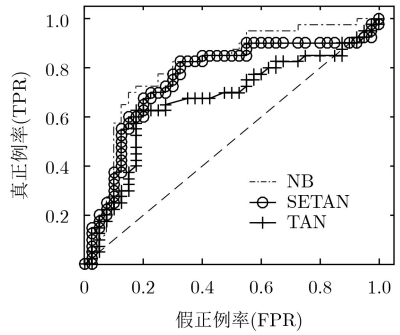


图 5 二分类数据集的ROC曲线

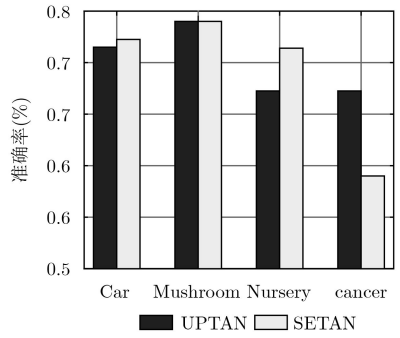


图 6 平均分类准确率比较图

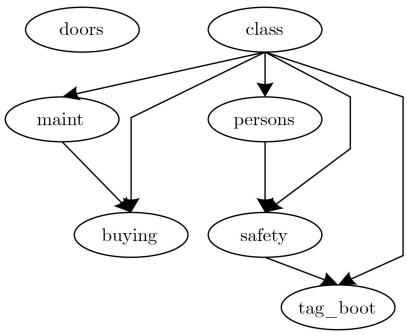


图 7 SETAN结构示意图

规模的数据集，但在SPECT和Connect数据集上的分类准确率较差，说明属性数量对分类模型的影响比较明显。其原因是，对于具有22个属性的SPECT数据集，80个样本相对于网络复杂度而言数据集规模太小，分类模型欠拟合导致各项分类指标不佳；对于Connect数据集，样本数量和属性数量均较大，相应的计算复杂度较高，导致评分搜索得到的模型指标不太理想。

总之，在数据规模、类别分布、属性数量这3个因素上，数据集的规模和类别分布对3种分类器的影响都比较小，而属性数量会明显影响分类效果。属性越多，分类准确率相应下降，但SETAN相比NB和TAN模型来说仍然占有优势。而且对于类别分布不均衡的数据集(如Balance, Car, Nursery)，SETAN的分类准确率有明显改善。

为了更直观地观察SETAN算法与TAN, NB算法的分类效果差异，图3给出了3种算法的AUC面积的polar图，图4为NB, TAN和SETAN 3种分类器在同一数据集上的分类准确率对比图。由于3种算法在Mushroom数据集上的AUC值非常接近，因此图3和图4没有给出。在图3中可以明显看出SETAN在各个polar图中面积都是最大的，说明SETAN模型整体上优于NB和TAN，图4的平均分类准确率则辅证了这一点；此外，SPECT属于二分类的小数据集，图5中给出了3种算法的ROC曲线图，可以看出，在处理属性数较多的小数据集SPECT时，SETAN算法的也具有很好的分类结果。

进一步验证SETAN树形分类器的有效性，与文献[4]的UPTAN分类器进行比较，图6是二者平均分类准确率的柱状图。由图可知，SETAN在Cancer数据集上差于UPTAN，这是由于Cancer数据集中属性值划分不一造成的，UPTAN所提出的不确定条件互信息度量能有效应对这一情况，而SETAN在Car, Mushroom, Nursery上均有良好表现。

图7是基于表3中数据集Car上的SETAN结构图，图7的class结点是类别结点，其余结点是与是否买车相关的属性结点。可以看出，本文提出的学习方法剔除了与买车不太相关的doors属性，在一定程度上进行了属性选择，学习到的SETAN模型不仅有良好的分类准确率，同时也构造了更加精简有效的树形结构。

4　结束语

本文提出一种基于评分搜索的树增强朴素贝叶斯分类器改进方法。考虑到属性对类贡献程度有所不同，该分类算法在此约束条件下利用低阶CI测试获得候选无效属性，随后通过改进的BIC评分函数结合K2算法的方式确定网络结构中弧边的方向，并去除无效属性，进而构建分类模型。本文方法额外允许属性没有父结点或只有一个属性父结点，从而构建了一种更好的树形贝叶斯网络结构，去除了冗余属性，增强了分类模型的可靠性。该算法和TAN分类模型的时间复杂度相同。实验结果表

明，与NB, TAN及UPTAN分类器相比，SETAN的分类准确率更高。下一步尝试在大规模数据集上进行该算法的分布式并行化研究。

参 考 文 献

[1] PEARL J. Probabilistic reasoning in intelligent systems: networks of plausible inference[J]. *Computer Science Artificial Intelligence*, 1991, 70(2): 1022–1027. doi: 10.2307/407557.

[2] WEBB G I, CHEN Shenglei, and N A. Zaidi Scalable learning of Bayesian network classifiers[J]. *Journal of Machine Learning Research*, 2016, 17(1): 1515–1549.

[3] MURALIDHARAN V and SUGUMARAN V. A comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis[J]. *Applied Soft Computing*, 2012, 12(8): 2023–2029. doi: 10.1016/j.asoc.2012.03.021.

[4] Friedman N, Geiger D, and Goldszmidt M. Bayesian network classifiers[J]. *Machine Learning*, 1997, 29(2-3): 131–163. doi: 10.1023/a:1007465528199.

[5] GAN Hongxiao, ZHANG Yang, and SONG Qun. Bayesian belief network for positive unlabeled learning with uncertainty[J]. *Pattern Recognition Letters*, 2017, 90. doi: 10.1016/j.patrec.2017.03.007.

[6] JIANG Liangxiao, CAI Zhihua, WANG Dianhong, *et al*. Improving Tree augmented Naive Bayes for class probability estimation[J]. *Knowledge-Based Systems*, 2012, 26: 239–245. doi: 10.1016/j.knosys.2011.08.010.

[7] 王中锋, 王志海. TAN分类器结构等价类空间及其在分类器学习算法中的应用[J]. 北京邮电大学学报, 2012, 35(1): 72–76. doi: 10.3969/j.issn.1007-5321.2012.01.017.

WANG Zhongfeng and WANG Zhihai. Equivalent classes of TAN classifier structure and their application on learning algorithm[J]. *Journal of Beijing University of Posts and Telecommunications*, 2012, 35(1): 72–76. doi: 10.3969/j.issn.1007-5321.2012.01.017.

[8] DUAN Zhiyi and WANG Limin. K-dependence bayesian classifier ensemble[J]. *Entropy*, 2017, 19(12): 651. doi: 10.3390/e19120651.

[9] 冯月进, 张凤斌. 最大相关最小冗余限定性贝叶斯网络分类器学习算法[J]. 重庆大学学报, 2014, 37(6): 71–77. doi: 10.11835/j.issn.1000-582X.2014.06.011.

FENG Yuejin and ZHANG Fengbi. Max-relevance min-redundancy restrictive BAN classifier learning algorithm[J]. *Journal of Chongqing University: Natural Science*, 2014, 37(6): 71–77. doi: 10.11835/j.issn.1000-582X.2014.06.011.

[10] WONG M L and LEUNG K S. An efficient data mining method for learning bayesian networks using an evolutionary algorithm-based hybrid approach[J]. *IEEE Transactions on Evolutionary Computation*, 2004, 8(4): 378–404. doi: 10.1109/TEVC.2004.830334.

[11] LOU Hua, WANG Limin, DUAN Dingbo, *et al*. RDE: A novel approach to improve the classification performance and expressivity of KDB[J]. *Plos One*, 2018, 13(7): e0199822. doi: 10.1371/journal.pone.0199822.

[12] ROBINSON R W. Counting Unlabeled Acyclic Digraphs[M]. Berlin Heidelberg: Springer, 1977: 28–43. doi: 10.1007/BFb0069178.

[13] SCHWARZ G. Estimating the Dimension of a Model[J]. *Annals of Statistics*, 1978, 6(2): 15–18.

[14] GREINER R and ZHOU W. Structural extension to logistic regression: discriminative parameter learning of belief net classifiers[J]. *Machine Learning*, 2005, 59(3): 297–322. doi: 10.1007/s10994-005-0469-0.

[15] MADDEN M G. On the classification performance of TAN and general bayesian networks[J]. *Knowledge-Based Systems*, 2009, 22(7): 489–495. doi: 10.1016/j.knosys.2008.10.006.

[16] DRUGAN M M and WIERING M A. Feature selection for Bayesian network classifiers using the MDL-FS score[J]. *International Journal of Approximate Reasoning*, 2010, 51(6): 695–717. doi: 10.1016/j.ijar.2010.02.001.

[17] WU Jiehua. A generalized tree augmented naive bayes link prediction model[J]. *Journal of Computational Science*, 2018. doi: 10.1016/j.jocs.2018.04.006.

[18] MEHRJOU A, HOSSEINI R, and ARAABI B N. Improved Bayesian information criterion for mixture model selection[J]. *Pattern Recognition Letters*, 2016, 69: 22–27. doi: 10.1016/j.patrec.2015.10.004.

[19] 杜瑞杰. 贝叶斯分类器及其应用研究[D]. [硕士论文], 上海大学, 2012.

DU Ruijie. The Research of Bayesian Classifier and its applications[D]. [Master disertation], Shanghai University, 2012

陈 曦：男，1963年生，教授，硕士生导师，研究方向为数据挖掘.
张 坤：男，1993年生，硕士生，研究方向为数据挖掘.