维普资讯 http://www.cqvip.com

第16卷 第5期
2006年5月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol. 16 No. 5
May 2006

TANC－BIC结构学习算法的改进

程泽凯，秦 锋，徐 浩
（安徽工业大学 计算机学院，安徽 马鞍山 243002）

摘 要：基于概率的贝叶斯分类器以其简单的结构和良好的性能受到重视，树扩展朴素贝叶斯分类器TANC应用较广。用TANC－BIC结构学习算法构建的分类器取得了成功，但TANC－BIC结构学习算法未考虑类节点的情况。文中提出了一种新的结构学习TANC－CBIC算法。并在贝叶斯分类器实验平台MBNC上编程实现。实验结果表明，改进算法分类准确率要高于由TANC－BIC和TANC－CMI结构学习算法构建的分类器，TANC－CBIC结构学习算法是有效的。

关键词：树扩展朴素贝叶斯分类器；贝叶斯信息标准测度；结构学习；数据采掘
中图分类号：TP301.6 **文献标识码**：A **文章编号**：1673－629X(2006)05－0044－03

Improvement for TANC－BIC Structure Learning Algorithm

CHENG Ze-kai, QIN Feng, XU Hao
(School of Computer Science, Anhui University of Technology, Maanshan 243002, China)

Abstract: Bayesian classifier based on probability theory has gained great attention, because of its simple structure and good performance. TANC applies widely in practice. The classifier which was set up by the TANC－BIC structure－learning algorithm had acquired success, but it didn't consider the class node. This paper suggests a new structure－learning algorithm called TANC－CBIC, makes experiment in MBNC experiment platform with programming TANC－CBIC algorithm. The results show that the accuracy of improver is better than algorithm based onTANC－BIC and TANC－CMI. The new structure－learning algorithm is effective.

Key words: TANC; BIC; structure－learning; data mining

0 前 言

在数据采掘技术中，分类能对大量的数据进行分析学习，并建立相应问题领域的分类模型。分类是数据采掘的研究热点，分类器的构建是分类的关键，基于概率的贝叶斯分类器以简单的结构和良好的性能受到重视。贝叶斯网络结构学习是建构分类器的关键，树扩展朴素贝叶斯TAN(Tree Augmented Naive Bayes)结构应用较广。

1 树扩展朴素贝叶斯分类器

贝叶斯网络BNs(Bayesian Networks)是结合图理论的概率模型，表达随机变量间的因果关系和概率关系。贝叶斯网络 $G=\langle S,P\rangle$ 是一个带有概率注释的有向无环图，由网络的拓扑结构 S 和局部概率分布的集合 P 两部分组成，S 中结点表示知识领域的随机变量，有向弧表示变量间的因果关系，P 是量化网络的一组参数，表达各结点间因果影响的强度。

贝叶斯网络分类器是一种特殊的贝叶斯网络，从已标签数据建立网络结构，再将未标签数据划分到类变量最大后验概率的那一类。完全的贝叶斯网络学习是NP难问题，TAN结构是贝叶斯网络的一种简化形式[1,2]。由TAN结构得到树扩展朴素贝叶斯分类器TANC(Tree Augmented Naive Bayesian Classifier)。

Nir Friedman[1]提出了TAN结构，属性变量以类变量作为父结点，属性结点间构成一棵树形结构，即 $pa(C)=\varnothing$，$C\in pa(A_i)$，$|pa(A_i)|<3$。如图1所示。

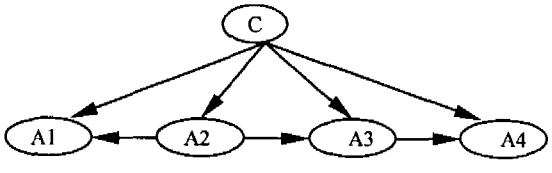

图1 树扩展朴素贝叶斯分类器模型

TANC的构建可分为网络结构学习和参数学习两部分，应用Bayes定理和条件独立假设：

$$C=\arg\max_{c_j\in C} p(c_j\mid a_1,\cdots,a_n)=\arg\max_{c_j\in C}\frac{p(a_1,\cdots,a_n\mid c_j)p(c_j)}{p(a_1,\cdots,a_n)}=\arg\max_{c_j\in C}p(a_1,\cdots,a_n\mid c_j)p(c_j)$$

类结点的最大后验概率 $C=\arg\max_{c_j\in C}p(c_j)\prod_{i=1}^{n}p(a_i\mid \Pi a_i)$，$C$ 表示TANC输出的目标值。

完整的TANC构造算法伪码简要叙述如下：

收稿日期：2005－09－08
基金项目：安徽省高等学校青年教师资助项目(2005JQ1079)
作者简介：程泽凯(1975－)，男，安徽马鞍山人，讲师，硕士，研究方向为人工智能、数据挖掘、机器学习；秦 锋，教授，研究方向为人工智能、数据挖掘、机器学习。

维普资讯 http://www.cqvip.com


```
VAR traindataset D; WeightMatrix Ip;
    UndirectedGraph UG; UndirectedTree UT;
    DirectedTree DT; DirectedGraph TAN;
    PriorType priors; Parameter CPT;
BEGIN
  for each Ai,Aj Compute Ip=f(Ai,Aj,D); %计算 i－j 结点对的关系矩阵 Ip
  G=ConstructUndirectedGraph(Ip); %建立非连接图
  UT=MaximumWeightedSpanningTree(G); %建立最大权重跨度树
  T=MakeDirected(UT); %确定边的方向
  TAN=AddClass(T); %添加类结点,得到 TAN 结构
  CPT=bayes_update_params(TAN,D,priors); %根据结构学习参数表
  TANC=(TAN,CPT); %创建贝叶斯分类器 TANC
  Clissfy testdataset; %分类测试例
  Evaluate Classifier; %评估分类器的性能
END
```

对于 $f(A_i,A_j,D)$ 的计算,Chow[3]采用的是互信息 MI(Mutual Information)测度,Friedman[1]使用的是条件互信息 CMI(Conditional Mutual Information)测度。由文献[1]可知,学习 TANC 结构的时间复杂性为 $O(n^2*N)$。n 是数据集中属性个数,N 是数据集中实例个数。并且,由于 CMI 测度考虑到了类结点与属性结点的关联,用 CMI 测度所构建的分类器的性能要优于用 MI 测度所构建的分类器。

2　基于贝叶斯信息标准的 TANC－BIC 结构学习算法

贝叶斯信息标准 BIC(Bayesian Information Criterion)是评估贝叶斯网络结构模型的尺度之一,在很好的拟和数据和模型的简洁性之间达到某种折衷。

其公式由 Schwarz[4]1978 年首先提出:给定数据集 D,$Q_{\mathrm{BIC}}(B,D)=LL(B\mid D)-\frac{1}{2}\log N*\mathrm{Dim}(B)$,其中,$LL(B\mid D)$ 是基于概率分布描述 D 所需要的比特数的度量,$\mathrm{Dim}(B)$ 是贝叶斯网络的维度,$\frac{1}{2}\log N$ 表示每一个参数使用的比特数。

笔者在文献[5]中提出了 TANC－BIC 结构学习算法。利用 BIC 测度函数计算 $i-j$ 属性对的相关性函数,TANC－BIC 结构学习算法在贝叶斯分类器实验平台 MBNC[6](Bayesian Networks Classifier using Matlab)上编程完成的,实验结果表明该算法是有效的。

3　TANC－BIC 结构学习算法的改进

TANC－BIC 结构学习算法假定实例的属性与类别是不相关的,未考虑类节点的情况,而在实际的数据集中,很多属性与类别是有关联的。这样,学习得到的结构不能在一定程度上影响到分类器的分类精度,具有一定的局限性。文中借鉴用 MI 测度与 TANC－CMI 测度学习贝叶斯分类器结构的思想,在 TANC－BIC 结构学习算法的基础上提出了 TANC－CBIC 结构学习算法。

TANC－CBIC 结构学习算法与 TANC－BIC 结构学习算法类似,主要区别在于对 $f(A_i,A_j,D)$ 的计算式进行改进。$f(A_i,A_j,D)=\mathrm{score}(A_i,A_j\cup C)-\mathrm{score}(A_i)$,$C$ 是类结点变量,A_i 和 A_j 是属性结点变量,score( ) 函数是用公式 $Q_{\mathrm{BIC}}(B,D)$ 对 $i-j$ 属性对打分的函数。BIC 的计算式 $Q_{\mathrm{BIC}}(B,D)$ 经过推导化简[7],可简化如式(1)所示:

$$Q_{\mathrm{BIC}}(B,D)=\sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}(N_{ijk}+\alpha_{ijk}-1)\log\frac{N_{ijk}+\alpha_{ijk}-1}{N_{ij}+\alpha_{ij}-r_i}-\frac{1}{2}\mathrm{Dim}(B)\log N \tag{1}$$

其中,X_i 有 r_i 个状态,其双亲集合(可能不止一个)有 q_i 个状态,$q_i=\prod_{x_p\in pa_i}r_p$,$N_{ijk}$ 是满足 X_i 的第 k 个状态,且 X_i 的双亲集合的第 j 个状态的记录数目;N_{ij} 是满足 X_i 的双亲的第 j 个状态的记录数目,即 $N_{ij}=\sum_{k=1}^{r_i}N_{ijk}$;$\alpha_{ijk}$ 和 α_{ij} 是先验信息,$\alpha_{ij}=\sum_{k=1}^{r_i}\alpha_{ijk}$。

TANC－BIC 结构学习算法仅仅考虑属性结点之间的相关性,而 TANC－CBIC 结构学习算法考虑了类结点和属性结点之间的相关性,理论上要优于 TANC－BIC 结构学习算法。

TANC－CBIC 结构学习算法的时复杂度也是 $O(n^2*N)$,属于多项式时间复杂度算法。与 TANC－BIC 结构学习算法的时间开销属同一数量级。

4　实验设计与结果分析

TANC－CBIC 结构学习算法是在 MBNC 实验平台上编程完成的。MBNC 实验平台可以方便地验证新的贝叶斯分类器家族学习算法,大大减少编程量,在 MBNC 上已经实现了用 TANC－CMI 结构学习算法和 TANC－BIC 结构学习算法学习得到的贝叶斯分类器。笔者对这两种分类器进行了比较。

文中使用 UCI[8](University of California in Irvine)的标准数据集作为主要实验数据来源,可保证与文献实验结果的可比性。有关经过预处理数据集的特征简要叙述见表 1。

分类器的建构分为结构学习和参数学习两部分。TANC－CBIC,TANC－BIC 和 TANC－CMI 三种分类器的结构学习算法不同,但参数学习算法是相同的。

准确性评估采用 5 叠交叉验证 CV5 和保留方法,数据集的划分方式与文献一致。准确性评估采用分类准确率指标,即所有分类正确的例数占整个测试集的比例。实验在同一环境下完成。实验结果见表 2。第 1 列是文中提出的算法得到的分类器的结果;第 2 列用 TANC－BIC 结构学习算法得到的分类器的实验结果;第 3 列是用 TANC－CMI 结构学习算法得到的分类的实验结果。

维普资讯 http://www.cqvip.com

表 1 数据集的简要概况

数据集	属性	类别	训练	测试	数据集	属性	类别	训练	测试
Australiaon	14	2	690	CV5	Lymphography	18	4	148	CV5
Breast	9	2	683	CV5	Monfn3-7-10	10	2	300	1024
Car	6	4	1880	CV5	Monk1	6	2	124	432
Chess	36	2	2130	1066	Monk2	6	2	169	432
Cleve	10	2	296	CV5	Monk3	6	2	122	432
Corral	6	2	128	CV5	Mushroom	22	2	9120	CV5
Crx	15	2	653	CV5	Nursery	8	2	11025	CV5
Diabetes	8	2	768	CV5	Pima	5	2	768	CV5
Flare	10	2	1066	CV5	Satimage	36	7	4435	2000
German	15	2	1000	CV5	Segment	18	7	1540	770
Glass	9	7	214	CV5	Shuttle-small	8	7	3866	1934
Glass2	9	2	163	CV5	Soybean	35	15	562	CV5
Heart	13	2	270	CV5	Vehicle	18	4	846	CV5
Hepatitis	19	2	80	CV5	Vote	16	2	435	CV5
Iris	4	3	150	CV5	Waveform21	18	3	300	4700
Letter	16	26	15000	5000					

表 2 测试实验结果 (%)

数据集\算法	TANC-CBIC	TANC-BIC	TANC-CMI	数据集\算法	TANC-CBIC	TANC-BIC	TANC-CMI
Australiaon	87.68	85.36	84.93	Lymphography	86.9	86.21	84.14
Breast	97.21	96.62	96.91	Monfn3-7-10	91.6	91.5	90.14
Car	94.49	92.87	91.28	Monk1	94.91	68.52	95.83
Chess	92.39	88.46	92.49	Monk2	65.28	67.36	68.98
Cleve	82.71	84.07	80.34	Monk3	96.06	95.6	97.22
Corral	98.4	84	99.2	Mushroom	100	99.97	99.93
Crx	88.31	87.69	85.39	Nursery	93.9	91.89	90.75
Diabetes	79.09	78.95	76.99	Pima	78.04	77.52	76.73
Flare	82.54	82.72	83.1	Satimage	87.25	85.45	86.8
German	74.2	73.5	72.8	Segment	92.34	93.77	94.81
Glass	69.05	75.71	68.57	Shuttle-small	99.43	99.48	99.48
Glass2	82.5	83.13	83.75	Soybean	94.46	89.82	91.96
Heart	84.81	83.7	83.7	Vehicle	72.9	72.54	72.66
Hepatitis	90	91.25	86.25	Vote	94.02	92.41	95.4
Iris	94	95.33	95.33	Waveform21	79.38	78.68	78.47
Letter	80.72	84.9	85.38	平均值(%)	87.24	85.77	86.76

从表中数据可初步得到:

(1)在计算结点对中考虑了类变量的情况,总体上分类效果明显增加,TANC-CBIC 结构学习算法比 TANC-BIC 结构学习算法的分类准确率要高;

(2)TANC-CBIC 结构学习算法比 MBNC 实验平台 TANC-CMI 结构学习算法的分类准确率要高些。

这表明文中提出的 TANC-CBIC 结构学习算法是有效的。

5 展 望

文中在 TANC-BIC 结构学习算法的基础上提出了一种新的 TANC-CBIC 结构学习算法,在 MBNC 实验平台上验证了新算法的有效性。进一步的研究如下:

(1)文中采用 BIC 测度计算 $f(A_i, A_j, D)$,以后尝试用更多的测度函数建构 TANC。

(2)最大跨权树根结点的选取,文献中没有涉及这方面的内容,一般是随机指定一个属性结点作为根节点的,具有很大的任意性。笔者认为:根结点的选取对分类效果还是有影响的,具体的还在进一步研究中。

参考文献:

[1] Friedman N. Bayesian network classifiers[J]. Machine Learning,1997 (29):131-163.

[2] 林士敏,田凤占,陆玉昌.用于数据采掘的贝叶斯分类器研究[J].计算机科学,2000, 27(10):73-76.

[3] Chow C K, Liu C N. Approximating Discrete Probability Distribution with Dependence Trees[J]. IEEE Trans on Information Theory,1968(14):462-467.

[4] Schwarz G. Estimating the dimension of a model. [J]. Annals of Statistics,1978,6,461-464.

[5] 程泽凯,林士敏. TANC-BIC 结构学习算法.[J].微机发展,2004,14(11):10-12.

[6] 程泽凯,林士敏,陆玉昌,等.基于 Matlab 的贝叶斯分类器平台 MBNC[J].复旦学报,2004,43(5):729-732.

[7] Sacha J P. New Synthesis of Bayesian Network Classifiers and Cardiac SPECT Image Interpretation[D]. USA: University of Toledo,1999.

[8] Blake C, Keogh E, Merz C. UCI repository of machine learning databasea[EB/OL]. http://www.ics.uci.edu/mlearn/MLRepository.html.1998.

(上接第 43 页)

[2] VAho A, Denning P, Ullman J. Principles of Optimal Page Replacement[J]. In Journal of ACM, 1971,18:80-93.

[3] Borodin A, Irani S, Raghavan P, et al. Competitive Paging with Locality of Reference[J]. In Journal of Computer and System Sciences,1995,50:244-258.

[4] Chrobak M, Noga J. LRU is Better than FIFO[A]. In Proc 9th Annual ACM-SIAM Symp on Discrete Algorithms[C]. Philadelphia, PA, USA: SIAM,1998.78-81.

[5] Juurlink B. Approximating the Optimal Replacement Algorithm[A]. In Proc 1st conference on computing frontiers[C]. New York, USA: Association for Computing Machinery, 2004.14-16.

[6] Irani S, Karlin A, Phillips S. Strongly Competitive Algorithms for Paging with Locality of Reference[A]. In 3rd Annual ACM-SIAM Symposium on Discrete Algorithms[C]. New York, NY, USA: ACM,1992.228-236.

[7] Glass G, Cao P. Adaptive Page Replacement Based on Memory Reference Behavior[A]. In ACMSIGMETRICS Conference on Measurement and Modeling of computer systems[C]. New York, NY, USA: ACM,1997.115-126.

欢迎订阅,欢迎刊登广告,电话:029-85522163