用于数据挖掘的TAN分类器的研究与应用

孙笑微,赵大宇,李晓毅,唐恒永
(沈阳师范大学,辽宁 沈阳 110034)

摘 要:分类是数据挖掘和模式识别中的一个基本和重要的课题。文中讨论了基于贝叶斯学习的TAN分类器的基本概念和分类算法,同时将分类器算法和具体分类算法结合为一个完整的有效算法。用某高校人才识别系统这一实例来说明TAN分类器的推理过程,并介绍了TAN分类器在数据挖掘领域的应用。实验结果表明TAN分类器具有较好的分类性能和较高的分类精度。

关键词:数据分类;TAN分类器;贝叶斯网络
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2006)11-0140-03

Study and Application of TAN Classifier for Data Mining

SUN Xiao-wei, ZHAO Da-yu, LI Xiao-yi, TANG Heng-yong
(Shenyang Normal University, Shenyang 110034, China)

Abstract: Classification is a basic and important task in data mining and pattern recognizing. In this paper, we discuss the basic concepts of TAN classifier and the algorithm based on Bayesian learning. Join the classifier algorithm and the concrete classification algorithm into an effective algorithm. The reasoning process of evaluating university talented scholars system is presented, and introduce the application of TAN classifier in data mining. The results prove that TAN classifier has perfect classification capability and higher classification accuracy.

Key words: data classification; TAN classifier; bayesian network

0 引 言

数据分类[1]是数据挖掘领域中一个非常重要的研究课题,其目的是分析输入数据,通过在训练集中的数据表现出来的特性,为每一个类找到一种准确的描述或模型,用由此生成的类描述对未来的测试数据进行分类。

进行分类首先要构造一个分类器。所谓分类器[2]是一个函数 $f(x)$,它给需要分类的实例 x 赋予类标签 $c_j \in C(j=1,2,\cdots,m)$,实例 x 由一组属性值 $X_1,X_2,\cdots,X_n$ 描述,C 是类变量,取有限个值,可看成有限个元素的集合。从预先分类的实例进行有导师学习并建立分类器是机器学习的中心问题之一。已有的分类器如决策树、决策表、神经网络、决策图和规则等,都可以看成不同的函数表示法。贝叶斯分类器指的是基于贝叶斯学习方法的分类器。用贝叶斯网络分类器进行分类的过程,实际上就是将属性结点作为证据结点引入到贝叶斯网络中,求得类结点后验概率的过程,后验概率最大时,类别结点相应的取值即作为分类的结果,这一过程又称为贝叶斯网络的推理[3]。

收稿日期:2006-02-26
基金项目:国家自然科学基金资助项目(10471096);知识科学与知识管理研究中心资助项目(027)
作者简介:孙笑微(1980-),女,辽宁本溪人,硕士研究生,研究方向为数据挖掘;赵大宇,博士,教授,研究方向为数据挖掘。

1 TAN分类器的概念及推理算法

1.1 TAN分类器的相关概念

TAN分类器[4]是由Nir Friedman等首次提出来的,是对朴素贝叶斯网络分类器的扩展。TAN分类器是树扩展朴素贝叶斯网络(tree-augmented naive Bayesian network)的简称,是以类变量为根结点,每个属性变量以类变量和最多一个属性变量为父结点的贝叶斯网络。

定义1 条件互信息 $I_{ij}=(X_i;X_j\mid C)=\sum_{X_i,X_j,C} p(X_i;X_j;C)\log_2\frac{p(X_i;X_j\mid C)}{p(X_i\mid C)p(X_j\mid C)}$ $i\neq j$

这里 X_i 与 X_j 为属性变量,C 为类变量。

定义2 将边按权重由大到小排序,遵照被选择的边不能构成回路的原则,按照边的权重由大到小的顺序选择边所构成的树称为最大权重跨度树。

定义3 类标签 $C_{NB}=\arg\max_{c_j\in C} P(c_j)\prod_{i=1}^{n} P(X_i\mid Pa_i,c_j)$,其中 X_i 为属性变量,Pa_i 为贝叶斯网络中除去类变量结点 C,结点 X_i 的父结点的集合。

学习TAN分类器是一个人机交互的过程,其方法基于1968年Chow和Liu提出的学习树结构的贝叶斯网络的方法。该方法将构造最大似然树的问题简化为在一个图中寻找最大权重跨度树的问题,使得所选择的弧构成一棵树,而且附属于选择弧的权重之和为最大。TAN分类器的推理一般分为以下3个步骤:(1)构建贝叶斯网络 S;

(2) 根据网络结构学习参数,即计算网络中各结点的局部条件概率 P;(3) 对由(S,P)确定的 TAN 分类器进行推理。

可以看出,构造贝叶斯网络是 TAN 分类器学习的关键所在,下面就给出 TAN 分类器推理算法。

1.2 TAN 分类器推理算法

具体算法描述如下:

输入:训练集和一组记录值 $A = \{a_1, a_2, \cdots, a_n\}$。

输出:类标签 C_{NB} 对应的最大概率 P。

算法:

1) 初始化图 $G(V,E)$, $V = \{$训练集中所有属性结点$\}$, $E = \varnothing$;

2) $S = \varnothing$;

3) for $i = 1$ to n //n 表示属性结点个数

4) for $j = 1$ to n

5) {

6) 计算条件互信息 $I_{ij}(i \neq j)$;

7) 将 I_{ij} 所对应的结点对放入 S 中;

8) }

9) 将 S 中的结点对按 I_{ij} 由大到小的顺序排列;

10) 从 S 中取出第一对结点,且把这对结点从 S 中删除;

11) 将从 S 中取出的结点对对应的边加到 E 中,构建最大权重跨度树;

12) 选择 V 中的一个点作为根结点构建有向树;

13) for $i = 1$ to n

14) 增加 C 作为每个 X_i 的父结点;

15) for $j = 1$ to m //计算类变量结点 C 的概率

16) $p_j = \frac{a_j}{a}$; //a_j 是类 c_j 中的训练样本数,a 是训练样本总数

17) for $j = 1$ to m

18) for $i = 1$ to n //计算属性结点 X_i 的条件概率

19) {

20) if X_i 是离散属性 $P_i = a_{ij}/a_j$; //a_{ij} 是在属性 X_i 上的值与父结点上的值分别对应且属于类 c_j 的训练样本数,a_j 是 c_j 中的训练样本数

21) if X_i 是连续属性 $P_i = \frac{1}{\sqrt{2\pi}\sigma_{C_j}} e^{-\frac{(X_i-\mu_{C_j})^2}{2\sigma_{C_j}^2}}$; //$\mu_{C_j}, \sigma_{C_j}$ 分别为高斯密度的平均值和标准差

22) }

23) $P = 0$;

24) for $j = 1$ to m //求属性结点 $X_1, X_2, \cdots, X_n$ 条件下 c_j 的概率

25) {

26) $Pa_0 = 1$;

27) for $i = 1$ to n

28) $Pa_i = Pa_{i-1} * Pa_i$; //Pa_i 为 X_i 的父结点的概率

29) $P_t = P_j * Pa_n$; //P_j 为属于类 c_j 的结点的概率

30) if $P_t > P$ $P = P_t$;

31) }

32) return P

2 TAN 分类器的应用

将上述 TAN 分类器推理算法部分应用于沈阳师范大学的人才评估系统中,根据该校现有的历史数据可以对人才进行评估考核。现以沈阳师范大学 1992 年至 2004 年间的 425 名教师为例,采用 TAN 分类器推理算法,对数据进行预处理后,除去数据中的冗余信息,有 5 个属性:其他状况(O),职称(T),学历(R),年龄(A)以及科研能力(S)。类别(C)是考核结果,分为优秀、合格、不合格三类。数据汇总一方面是将各院系的数据进行集成;另一方面是将数据进行概化处理,即将低层次的原始数据替换为高层次的概念,以便于进行数据挖掘。例如将具体的年龄概化为≤30,30~50 和>50 三个年龄段,分别代表青年、中年和老年教师。这里教师的科研能力只用论文的数量来表示(其实科研能力还应包括出版专著及申请项目,这里忽略不计),概化为≤5,5~20 和>20 三档,分别代表科研能力差、良、优。具体数据整理成表 1,其包含 425 条相关记录。由于篇幅关系,文中只列出原表的基本结构和表中的部分数据。

表 1　高校人才评估历史汇总数据表

教师号	其他状况(O)	职称(T)	学历(R)	年龄(A)	科研能力(S)	考核结果(C)
001	院士	教授	博士	>50	>20	优秀
002	理事长	副教授	硕士	>50	5~20	优秀
003	兼职教授	副教授	硕士	30~50	5~20	合格
…	…	…	…	…	…	…
425	其他	讲师	学士	30~50	5~20	不合格

(1)设计网络结构 S。

计算任意两个属性结点的条件互信息值并按大小排序得 $I(T;A \mid C) > I(R;A \mid C) > I(A;S \mid C) > I(O;A \mid C) > I(T;S \mid C) > I(O;T \mid C) > I(O;S \mid C) > I(O;R \mid C) > I(T;R \mid C) > I(R;S \mid C)$。由此构建最大权重跨度树(见图 1)和贝叶斯网络结构图(见图 2)。

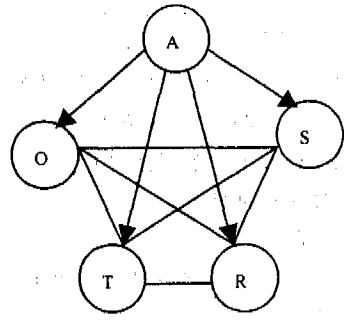

图 1　最大权重跨度树(带箭头)

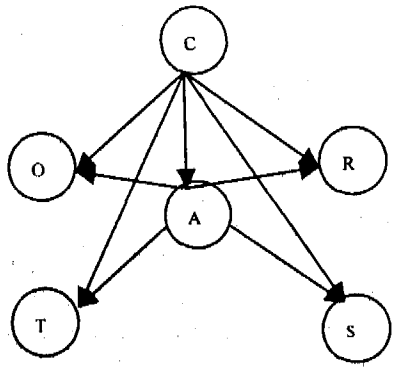

图 2 贝叶斯网络结构图

(2)求各属性结点概率。

设计网络时,必须输入某一态势的概率。也就是说,专家指定每个结点的条件概率表。本例中有 $P(C)$,$P(A \mid C)$,$P(O \mid C,A)$,$P(T \mid C,A)$,$P(R \mid C,A)$,$P(S \mid C,A)$。

条件概率表由表 1 计频数给出,如图 3 所示,其中 0.5/N 表示数据库中为 0 的记录(N 为训练样本总数)。

C	
优秀	0.5
合格	0.15
不合格	0.35

A	优秀	合格	不合格
>50	0.5	0.333	0.286
30~50	0.4	0.333	0.571
≤30	0.1	0.333	0.143

O	优秀			合格			不合格		
	>50	30~50	≤30	>50	30~50	≤30	>50	30~50	≤30
院士	0.6	0.25	0.5/N	0.5/N	0.5/N	0.5/N	0.5/N	0.5/N	0.5/N
兼职	0.2	0.5	0.5/N	0.5	0.5	0.5/N	0.5	0.4	0.5/N
其他	0.2	0.25	1	0.5	0.5	1	0.5	0.6	1

T	优秀			合格			不合格		
	>50	30~50	≤30	>50	30~50	≤30	>50	30~50	≤30
教授	0.4	0.75	0.5/N	0.2	0.1	0.5/N	0.5	0.1	0.5/N
副教授	0.6	0.25	0.5/N	0.7	0.8	0.5/N	0.5	0.15	0.5/N
讲师	0.5/N	0.5/N	1	0.1	0.1	1	0.5/N	0.75	1

R	优秀			合格			不合格		
	>50	30~50	≤30	>50	30~50	≤30	>50	30~50	≤30
博士	0.4	0.75	0.5/N	0.5	0.5	0.1	0.2	0.5/N	0.5/N
硕士	0.6	0.25	1	0.5	0.5	0.7	0.7	0.75	0.2
学士	0.5/N	0.5/N	0.5/N	0.5/N	0.5/N	0.2	0.1	0.25	0.8

S	优秀			合格			不合格		
	>50	30~50	≤30	>50	30~50	≤30	>50	30~50	≤30
>20	0.4	0.25	0.5/N	0.5	0.2	0.5/N	0.5/N	0.5/N	0.5/N
5~20	0.6	0.6	0.95	0.5	0.8	0.5/N	0.5	0.5	0.5/N
≤5	0.5/N	0.15	0.05	0.5/N	0.5/N	1	0.5	0.5	1

图 3 贝叶斯网络条件概率表

图 2 和图 3 构成了一个完整的贝叶斯网络,此网络可看作一个分类器进行推理。

(3)推理过程。

如果高校决策者希望通过上面的贝叶斯网络,预测 30~50 岁之间科研能力很强、拥有博士学位且作为兼职教授的中年副教授的评估结果,即未知样本 X = (O = "兼职教授",T = "副教授",R = "博士",A = "30 ~ 50",S = ">20")。

计算 $C_{NB} = P(C) * P(A \mid C) * P(O \mid C;A) * P(T \mid C;A) * P(R \mid C;A) * P(S \mid C;A)$

则 C(优秀) = $0.5 * 0.4 * 0.5 * 0.25 * 0.75 * 0.25 = 4.69 * 10^{-3}$

C(合格) = $0.15 * 0.333 * 0.5 * 0.8 * 0.5 * 0.2 = 2 * 10^{-3}$

C(不合格) = $0.35 * 0.571 * 0.4 * 0.15 * 0.001176 * 0.001176 = 1.66 * 10^{-8}$

从以上数据的比较可以知道此样本的分类结果为优秀。通过这些数据和 TAN 分类器,可以帮助学校领导做出提高整体科研水平的决策:如高校人才引进的重点应为两院的院士,在国内学术界有较大影响的学科带头人其中包括在国家高级学会担任过重要职务的人员,以及 50 岁以下具有博士学位科研能力很强的教授和 30 岁以下有较强科研能力的年轻硕士。

从这个例子中可以看到 TAN 分类器推理算法很好地反映了各属性变量与类变量之间的关系,为高校人才的考核提供了科学依据。

3 结束语

用 TAN 分类器找出数据间的潜在关系,正是数据挖掘所要完成的功能[4]。但是 TAN 分类器只适用于少量数据的数据库中,当数据库中的数据量十分庞大时,TAN 分类器效果不如 BAN 和 GBN 分类器效果好[5]。在实际应用大规模数据库的时候,贝叶斯分类器还有很多待研究的问题。

参考文献:

[1] 邵峰晶,于忠清.数据挖掘原理与算法[M].北京:中国水利水电出版社,2003.

[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] 余东峰,孙兆林.基于贝叶斯网络不确定推理的研究[J].微型电脑应用,2004,20(8):6-8.

[4] Russell S, Binder J, Koller D, et al. Local learning in probabilistic networks with hidden variables[C]// In: Cooper G F, Moral S ed. Proceedings of the 14th International Joint Conference on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers, Inc.1998:1146-1152.

[5] Cheng J, Greiner R. Comparing Bayesian network classifiers [C]// Proceedings of the fifteenth conference on uncertainty in artificial intelligence. San Francisco: Morgan Kaufmann. 1999:101-107.