文章编号:1671-8860(2007)04-0287-03　　文献标志码:A

基于 Tree Augmented Naive Bayes Classifier 的影像纹理分类

虞　欣[1]　郑肇葆[1]　叶志伟[1]　田礼乔[2]

(1　武汉大学遥感信息工程学院,武汉市珞喻路 129 号,430079)
(2　武汉大学测绘遥感信息工程国家重点实验室,武汉市珞喻路 129 号,430079)

摘　要:提出了一种松弛方法,允许类别节点下的相邻子节点之间存在相关关系(有向边),这种方法称为树增强型简单贝叶斯分类器(tree augmented naive Bayes classifier, TAN)。实验结果表明,TAN 比简单贝叶斯分类器(naive Bayes classifier,NBC)可以获得更高的分类精度。

关键词:贝叶斯网络;纹理分类;影像
中图法分类号:TP753; P237.4

1988 年,Pearl 等人就提出了贝叶斯网络的概念[1]。作为一种不确定性知识表达与推理工具,贝叶斯网络在医疗诊断、故障检测和软件测试等方面有着广泛的应用[2]。然而在分类领域中,直到简单贝叶斯网络(naive Bayes network)[3]的出现,它才引起研究工作者的关注。在简单贝叶斯网络模型中,要求(假设)父节点下的子节点之间相互保持独立,但这在现实世界中往往并非如此。鉴于此,本文提出一种对"天真"假设的松弛方法,即允许父节点下的子节点之间有相关关系(存在有向边),但考虑到计算工作量,不允许这些子节点之间可以有任意的有向边,而只是把这些子节点之间的关系限制为一种树型关系,称为树增强型简单贝叶斯网络,它在分类领域中习惯称为树增强型简单贝叶斯分类器(TAN)。

1　基于 TAN 的分类法

1.1　树增强型简单贝叶斯分类器(TAN)

与简单贝叶斯分类器(NBC)不同,TAN 允许类别节点下的相邻子节点之间存在相关关系。这种变量之间的相互关系(依赖性)的描述方法比 NBC 更接近现实世界的真实情况。

图 1 为 TAN 在分类中应用的一个图例,图中,C 为类别节点(变量),变量 X_i 为从某一个待分类单元中提取的纹理特征。根据贝叶斯网络[4]的定义有 $P_a(C)=\varnothing$,$P_a(X_1)=C$,$P_a(X_2)=\{C,X_1,X_3\}$。所有节点(事件)的联合概率为:

$$P(X_1,X_2,\cdots,X_n)=\prod_{i=1}^{n}P(X_i/P_a(X_i)) \quad (1)$$

实际上,从网络拓扑结构方面来看,TAN 是对 NBC 模型的一种推广,即先建立一个 NBC 模型,然后在它的基础上再考虑类别节点下的相邻子节点之间的相互关系,建立类似图 1 的一种网络拓扑结构,所以,TAN 比 NBC 多了一个网络拓扑结构的学习(或训练)阶段。然而要确定 TAN 的网络拓扑结构,必须根据某种准则或通过训练样本进行学习得到。

1.2　纹理特征的提取

纹理在影像分析与理解中一直是一个非常重要的内容。在本文的实验中,提取了7种纹理特

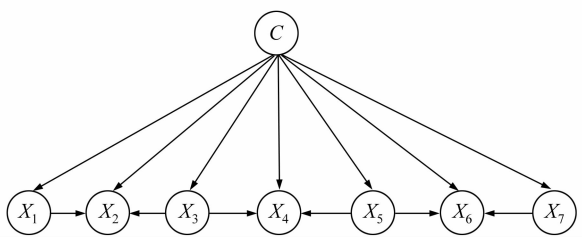

图 1　TAN 在分类中的应用
Fig. 1　TAN Applied in Classification

收稿日期:2007-01-20。
项目来源:国家自然科学基金资助项目(40271094,40571102)。

征，大致把它们分为两类：统计性纹理特征和结构性纹理特征。

1）偏度、信息熵和灰度共生矩阵中的逆差矩分别记为 X_1、X_2 和 X_3；

2）图像经过 Symlets 小波变换后，分别提取一尺度的近似分量 LL 中的均值、水平细节分量 LH 中的方差和垂直细节分量 HL 中的方差。另外，再加上分形维特征，分别记为 X_4、X_5、X_6 和 X_7。

1.3 网络拓扑结构和参数的学习

贝叶斯网络的学习（或训练）阶段包括两个部分，即网络拓扑结构的学习[5]和参数的学习（估计）[6]。其中，网络拓扑结构的学习一直是贝叶斯网络研究的热点，也是一个难点，特别是当节点个数较多的情况[7]。由于应用专业背景的不同，研究工作者提出了评多网络拓扑结构的学习方法，主要可以分为两大类，一类是基于评分函数的方法，如最小长度描述法；另一类是基于依赖关系分析的方法，如条件独立测试法。

然而，一方面，由于不像在医疗诊断应用中那样，节点之间存在一种明显的因果关系，如感冒（节点）会引起（指向）发烧（节点），在分类应用中，所提取的特征变量（节点）之间并不具备这种明显的因果关系；另一方面，考虑到计算工作量的问题，因此，在 TAN 模型中，只允许（假设）相邻的两个变量之间存在相关关系（有向边），即要么 $X_i \rightarrow X_{i+1}$，要么 $X_{i+1} \rightarrow X_i$，这样，特征变量之间形成的拓扑结构在形状上与“树”非常相像，通常把这个约束（假设）称为树型结构约束。在这个条件的约束下，对所有的拓扑结构进行穷尽搜索，还有 $n! \times 2^{n-1}$ 种。当 $n=7$ 时，有 322 560 种。如果假定一个特征变量的排列顺序有 $7! = 5\ 040$ 种，那么可能的网络拓扑结构就减少为 64 种。考虑到计算量的原因，在本文实验中，事先假定了一个特征变量的出现顺序为偏度、信息熵、灰度共生矩阵中的逆差矩、基于小波的结构特征和分形维特征。

本文提出一种根据训练样本的训练精度来学习网络拓扑结构的方法。首先，任意给定一种初始的状态（类似图 1），进行网络参数的学习；然后把训练样本当作测试样本进行分类，得到训练样本的分类精度（这个精度称为训练精度）；最后在树型结构的约束和顺序假设下，依次搜索所有可能的网络拓扑结构，其中最高的训练精度所对应的那个网络（包括拓扑结构和参数的估计值）作为学习的最好结果，用于后续的分类中。

1.4 概率推理模型

从图 1 可以看到，除类别节点 C 外，其他的节点至少有一个父节点，还可能存在其他的父节点，这种情况下的概率推理模型就比 NBC 要复杂一些。下面给出大致的推导过程[8]。

设 $\boldsymbol{X}_s$ 为 TAN 中的任意一个节点，记 $\boldsymbol{X}_p$ 为子节点 $\boldsymbol{X}_s$ 的父节点集（即 $P_a(\boldsymbol{X}_s)$）。在下面的表述中，用大写的英文字母表示随机变量，相应的小写字母表示该随机变量的观测值，也即样本值。假设 $\boldsymbol{X}_s$ 和 $\boldsymbol{X}_p$ 服从正态分布，两者可以组成一个 $(n+1)$ 维的正态随机向量 $\boldsymbol{X}$，即

$$\underset{(n+1)\times 1}{\boldsymbol{X}^{\mathrm{T}}} = [\underset{1\times 1}{\boldsymbol{X}_s} \quad \underset{n\times 1}{\boldsymbol{X}_p}] \tag{2}$$

则相应的均值向量 $\boldsymbol{\mu}_X$ 和协方差矩阵 $\boldsymbol{D}_X$ 为：

$$\boldsymbol{\mu}_X = \begin{bmatrix}\boldsymbol{\mu}_s \\ \boldsymbol{\mu}_p\end{bmatrix} = \overline{\boldsymbol{X}}, \boldsymbol{D}_X = \begin{bmatrix}\boldsymbol{D}_{ss} & \boldsymbol{D}_{sp} \\ \boldsymbol{D}_{ps} & \boldsymbol{D}_{pp}\end{bmatrix} = \frac{1}{n}(\boldsymbol{X}-\boldsymbol{\mu}_X)(\boldsymbol{X}-\boldsymbol{\mu}_X)^{\mathrm{T}} \tag{3}$$

从而有：

$$\boldsymbol{X}_s \sim N(\boldsymbol{\mu}_s, \boldsymbol{D}_{ss}), \boldsymbol{X}_p \sim N(\boldsymbol{\mu}_p, \boldsymbol{D}_{pp}) \tag{4}$$

所以 $\boldsymbol{X}$ 的概率密度为：

$$f(\boldsymbol{X}) = (2\pi)^{-\frac{n+1}{2}} \times |\boldsymbol{D}_X|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}\begin{bmatrix}\boldsymbol{X}_s-\boldsymbol{\mu}_s \\ \boldsymbol{X}_p-\boldsymbol{\mu}_p\end{bmatrix}^{\mathrm{T}} \boldsymbol{D}_X^{-1} \begin{bmatrix}\boldsymbol{X}_s-\boldsymbol{\mu}_s \\ \boldsymbol{X}_p-\boldsymbol{\mu}_p\end{bmatrix}\right\} \tag{5}$$

进而根据条件概率密度公式可以得到 $\boldsymbol{X}_s$ 对 $\boldsymbol{x}_p$ 的条件概率密度为：

$$f(\boldsymbol{X}_s \mid \boldsymbol{x}_p) = (2\pi)^{-\frac{n+1}{2}} \times |\widetilde{\boldsymbol{D}}_{ss}|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}[\boldsymbol{X}_s-\widetilde{\boldsymbol{\mu}}_s]^{\mathrm{T}} \widetilde{\boldsymbol{D}}_{ss}^{-1}[\boldsymbol{X}_s-\widetilde{\boldsymbol{\mu}}_s]\right\} \tag{6}$$

式(6)仍然是正态概率密度。式中，$\widetilde{\boldsymbol{\mu}}_s$ 为 $\boldsymbol{X}_s$ 对 $\boldsymbol{x}_p$ 的条件期望；$\widetilde{\boldsymbol{D}}_{ss}$ 为 $\boldsymbol{X}_s$ 对 $\boldsymbol{x}_p$ 的条件方差，可由下式计算：

$$E(\boldsymbol{X}_s \mid \boldsymbol{x}_p) = \widetilde{\boldsymbol{\mu}}_s = \boldsymbol{\mu}_s + \boldsymbol{D}_{sp}\boldsymbol{D}_{pp}^{-1}(\boldsymbol{x}_p - \boldsymbol{u}_p) \tag{7}$$

$$D(\boldsymbol{X}_s \mid \boldsymbol{x}_p) = \widetilde{\boldsymbol{D}}_{ss} = \boldsymbol{D}_{ss} - \boldsymbol{D}_{sp}\boldsymbol{D}_{pp}^{-1}\boldsymbol{D}_{ps} \tag{8}$$

最后可以计算出当 $\boldsymbol{X}_p$ 的观测值为 $\boldsymbol{x}_p$ 时 $\boldsymbol{X}_s$ 的条件概率：

$$P(\boldsymbol{X}_s \mid \boldsymbol{x}_p) = f(\boldsymbol{X}_s \mid \boldsymbol{x}_p)\mathrm{d}\boldsymbol{X}_s \tag{9}$$

式中，$\mathrm{d}\boldsymbol{X}_s$ 表示微小的变化量（步长），在计算中可视为 1。

根据式(1)和式(9)，由下式计算所有节点的联合概率为：

$$P(X_1, X_2, \cdots, X_n, C_i) = \prod_{s=1}^{n} P(\boldsymbol{X}_s \mid \boldsymbol{x}_p) \tag{10}$$

式中，C_i 为类别变量，$i=1, \cdots, m$ 表示相应的类别，m 为类别的总数。由条件概率可以得：

$$P(C_i \mid X_1, X_2, \cdots, X_n) = \frac{P(X_1, X_2, \cdots, X_n \mid C_i)P(C_i)}{P(X_1, X_2, \cdots, X_n)} = \frac{P(X_1, X_2, \cdots, X_n, C_i)}{P(X_1, X_2, \cdots, X_n)} \tag{11}$$

由于 $P(X_1,X_2,\cdots,X_n)$是一个与 C_i 无关的常量,所以有:

$$P(C_i \mid X_1,X_2,\cdots,X_n) \propto P(X_1,X_2,\cdots,X_n,C_i) \tag{12}$$

然后根据最大后验概率最大的准则进行判别,即 C^* 为$\max_i\{P(C_i \mid X_1,X_2,\cdots,X_n)\}$。

2　实验结果与分析

为了验证 TAN 模型在分类应用中的正确性和有效性,本文选取了澳大利亚某地区的 6 幅 23 cm×23 cm 的黑白航空影像和 10 幅武汉地区的 23 cm×23 cm 的黑白航空影像,根据野外调绘的结果,对这 16 幅大的航空影像人工分割为小块的 465 幅小图像,并将它们分成三类:居民地(167 幅)、田地(144 幅)和水域(154 幅),其中最小的为 16 像素×16 像素,最大的为 40 像素×40 像素。

图 2 为在每一类中随机地选取 50 个样本时,经过搜索后得到训练精度最高(90.67%)的网络拓扑结构(次优训练精度为 90%,此时的网络拓扑结构只是把图 2 中的 $X_4 \rightarrow X_5$ 改为 $X_4 \leftarrow X_5$)。

在图 3 中,横坐标表示从每一类中选取训练样本的个数,纵坐标表示相应的总的分类精度。为了验证 TAN 方法的有效性,与 NBC 和 PCA-NBC[1,9](NBC 的一种改进方法)进行了比较。从图 3 可以看到,基于 TAN 的总的分类精度较高,它的曲线在 PCA-NBC 和 NBC 曲线的上面,具体数据见表 1。

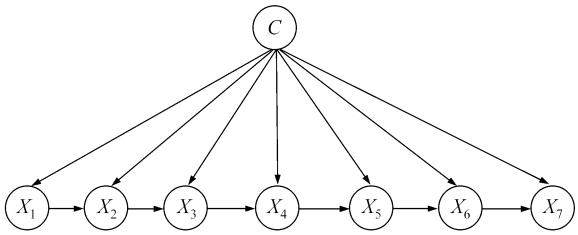

图 2　训练的 TAN 的拓扑结构

Fig. 2　Topology Structure of TAN

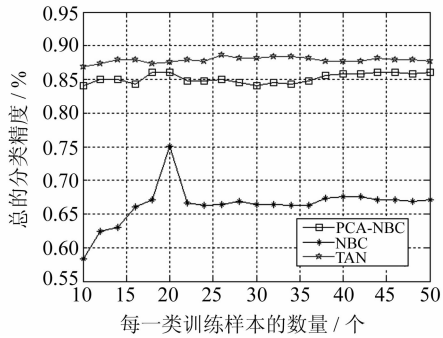

图 3　训练样本对总的分类精度的影响曲线

Fig. 3　Curve of Total Classification Accuracy

表 1　三种方法的比较

Tab. 1　Comparison of Three Methods

N	10	15	20	25	30	35	40	45	50	标准差
TAN	86.88	88.39	87.53	87.74	88.17	88.39	87.74	87.96	87.74	0.004 0
NBC	58.28	64.52	75.05	66.45	66.45	66.24	67.50	67.10	67.10	0.029 5
PCA-NBC	84.09	84.30	86.02	84.52	84.09	84.30	85.81	86.02	86.02	0.007 2

从表 1 可以看到,基于 TAN 的分类精度在[86.88%,88.39%]之间,而且标准差仅为0.004。这从另一个角度说明,训练样本的数量对总的分类精度的影响较小,分类精度比较稳定,而且令人满意。另外,PCA-NBC 方法的分类精度比 TAN 方法平均低 2.74%。

在今后的研究工作中,还需要对树增强型简单贝叶斯分类器进行更深入的研究,如特征变量出现的顺序对分类的影响,如何解释特征变量之间的指向关系所表示的内在的"物理"意义等。

参　考　文　献

[1] Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers[J]. Machine Learning, 1997(29):131-163

[2] Bart B, Geert V. Bayesian Network Classifiers for Identifying the Slope of the Customer Lifecycle of Long-life Customers[J]. European Journal of Operational Research, 2004(156):508-523

[3] Yu Xin, Zheng Zhaobao, Li Linyi, et al. Texture Classification of Aerial Image Based on PCA-NBC [C]. The 4th International Symposium on MIPPR, Wuhan, 2005

[4] Heckerman D, Geiger D, Chickering D M. Learning Bayesian Networks: the Combination of Knowledge and Statistical Data[J]. Machine Learning, 1995(20):197-244

[5] Cheng Jie, Greiner R. Learning Bayesian Belief Network Classifiers: Algorithms and System[C]. The 14th Canadian Conference on Artificial Intelligence, Canadian, 2001

(下转第 365 页)

ETL of Spatial Data Warehouse

TIAN Yangge[1]　*BIAN Fuling*[1]

(1　Research Center of Spatial Information and Digital Engineering, International Software Institute, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

Abstract: The data warehouse and ETL are introduced, the basic frame of the ETL of the spatial data warehouse is discussed, the ETL of Guangzhou agricultural economy data warehouse is presented as an example.

Key words: data warehouse; spatial data warehouse; ETL

About the first author: TIAN Yangge, Ph.D candidate, majors in GIS and data mining.

E-mail: tiandebox@126.com

(上接第 289 页)

[6] Cheng Jie, Greiner R. Comparing Bayesian Network Classifiers[C]. The 15th Conference on Uncertainty in Artificial Intelligence, San Francisco, 1999

[7] Yang Shulin, Chang Kuochu. Comparison of Score Metrics for Bayesian Network Learning[J]. IEEE Transactions on Systems, Man and Cybernetics—Part A: Systems and Humans, 2002, 32(3): 419-428

[8] 崔希璋,於宗俦,陶本藻,等. 广义测量平差[M]. 武汉:武汉测绘科技大学出版社,2001

[9] 虞欣,郑肇葆,汤凌,等.基于 Naive Bayes Classifiers 的航空影像纹理分类[J].武汉大学学报·信息科学版, 2006,31(2): 108-111

第一作者简介:虞欣,博士生。现主要从事图像解译和贝叶斯统计研究。

E-mail:china_yuxin@yahoo.com.cn

Texture Classification Based on Tree Augmented Naive Bayes Classifier

YU Xin[1]　*ZHENG Zhaobao*[1]　*YE Zhiwei*[1]　*TIAN Liqiao*[2]

(1　School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

(2　State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

Abstract: On the basis of the study of naive Bayes classifiers (NBC), a new method— tree augmented naive Bayes classifier is proposed and applied to texture classification. The experiment results demonstrate that the new method performs better in overall classification precision than NBC.

Key words: Bayesian network; texture classification; image

About the first author: YU Xin, Ph.D candidate, majors in the image interpretation and Bayesian statistics.

E-mail: china_yuxin@yahoo.com.cn