文章编号:1671-8860(2010)02-0228-05　　**文献标志码**:A

利用互信息改进遥感影像朴素贝叶斯网络分类器

陶建斌[1]　舒　宁[1]　沈照庆[1]

(1　武汉大学遥感信息工程学院,武汉市珞喻路 129 号,430079)

摘　要:针对朴素贝叶斯网络简单条件独立性假设的不足,将它的一种改进形式——选择型朴素贝叶斯网络和两种扩展形式(树增强型朴素贝叶斯网络、贝叶斯增强型朴素贝叶斯网络)用于多光谱遥感影像的分类中。在分析波段间互信息的基础上,分别构造了上述 3 种分类器,并和朴素贝叶斯网络分类器的性能进行了比较。

关键词:互信息;条件独立性分析;贝叶斯网络分类器;遥感影像;分类

中图法分类号:P237.3

朴素贝叶斯网络分类器(naïve Baysian classifier,NBC)以其简单的网络结构和良好的学习效率在许多领域都获得了成功[1-3],然而它的特征间相互独立的假设过于理想化[4-6]。近年来,大多数针对 NBC 的研究都是围绕着对特征进行数学变换,以满足条件独立性要求[5]或放松条件独立性假设[2,4,7]。针对遥感数据的特点,本文试图从不同角度放松这种假设,以期提高分类器的性能。根据互信息分析去掉部分相关性较强的特征,在减少分类特征的同时提高了分类器的性能,这便是选择型朴素贝叶斯网络分类器(selective naïve Bayesian classifier,SNBC)。如果将互信息融入到结构学习中,采用不同的结构学习算法,即可以分别得到树增强型朴素贝叶斯网络(tree augmented naïve Bayes,TAN)和贝叶斯增强型朴素贝叶斯网络(Bayes augmented naïve Bayes,BAN)。本文构造了上述 3 种分类器,并实现了对 TM 影像的分类。

1　贝叶斯网络分类器

贝叶斯网络是概率论与图论相结合的产物,它以图(graph)的形式描述节点(或变量)之间的关系,用概率表示变量间的相关关系或依赖程度。

构造贝叶斯网络分类器包括贝叶斯学习(主要是结构学习和参数学习)和后验概率推理两个阶段。首先根据先验信息获得网络结构,或从训练数据中学习得到网络结构;然后根据训练样本和网络结构通过概率统计的方法计算得出条件概率表;最后由条件概率表和待分类数据(测试样本)完成分类,其过程如图 1 所示。

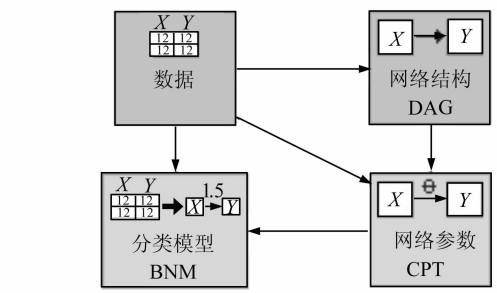


图 1　贝叶斯网络分类器模型示意图

Fig. 1　Schematic Diagram of Bayesian Network Classifier Model

分类器将分类特征向量作为输入,类后验概率作为输出,网络节点的取值对应变量集 $G=\{X_1,X_2,\cdots,X_n,C\}$,其中,$X_i(i=1,2,\cdots,n,n$ 为波段数)是遥感数据的若干个波段,特征的一个实例可用向量$(x_1,x_2,\cdots,x_n)$表示;C 表示类别变量,c_k 表示 C 的值,$k=1,2,\cdots,m,m$ 为类别数。则类后验概率可由贝叶斯公式计算得到:

$$p(c_k \mid x_1,x_2,\cdots,x_n)=\frac{p(x_1,x_2,\cdots,x_n \mid c_k)p(c_k)}{p(x_1,x_2,\cdots,x_n)}=\frac{p(x_1,x_2,\cdots,x_n,c_k)}{p(x_1,x_2,\cdots,x_n)} \tag{1}$$

式中,$p(c_k \mid x_1,x_2,\cdots,x_n)$是类 c_k 的后验概率;

收稿日期:2009-12-17。

项目来源:国家 973 计划资助项目(2006CB701303)。

$p(c_k)$是类 c_k的先验概率;$p(x_1,x_2,\cdots,x_n|c_k)$是类 c_k的似然度。$p(x_1,x_2,\cdots,x_n)$对各个类别都是常数,故有:

$$p(c_k \mid x_1,x_2,\cdots,x_n) \propto p(x_1,x_2,\cdots,x_n,c_k) \tag{2}$$

$p(x_1,x_2,\cdots,x_n,c_k)$是特征节点和类别节点的联合概率。

可见,计算后验概率的关键就是如何根据网络结构和每个节点的条件概率表计算得出这个联合概率。故根据链规则,联合概率可写为:

$$p(x_1,x_2,\cdots x_n,c_k) = p(c_k)\prod_{i=1}^{n} p(x_i \mid \pi(x_i)) \tag{3}$$

式(3)将联合概率分解为各节点的条件概率的乘积。接下来的关键是通过结构学习找到最能与样本数据拟合的网络结构,也就是找到每个特征节点的父节点(集)。

分类器根据后验概率最大的准则进行分类,即将类 c_k赋值为 $\max\{p(c_k|x_1,x_2,\cdots,x_n)\}$。

结构学习是贝叶斯学习的主要任务,本文采用从样本数据中学习得到网络结构的方法。

2　实验与分析

朴素贝叶斯分类器(NBC)将特征节点的父节点约束为类节点,而特征节点之间没有连接(见图 2,以 TM 影像的 1、2、3、4、5、7 波段为例)。这样,相比于式(3),联合概率的计算得到了极大的简化,即

$$p(x_1,x_2,\cdots x_n,c_k) = p(c_k)\prod_{i=1}^{n} p(x_i \mid c_k) \tag{4}$$

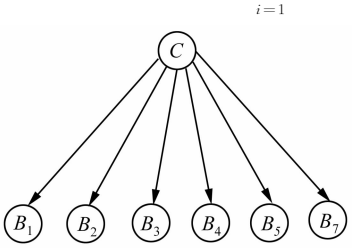


图 2　NBC 的网络结构

Fig. 2　Network Structure of NBC

NBC 假定各特征节点间是相互独立的,也就是说,各特征独立地作用于决策变量,而不考虑它们之间的相关性。然而,现实中,TM 影像的波段之间都有不同程度的相关性。

信息论中,节点间的互信息反映了两个变量是否相关及其相关性的程度。两个节点间的互信息定义为[8]:

$$I(X_i,X_j) = \sum_{x_i,x_j} p(x_i,x_j)\lg\frac{p(x_i,x_j)}{p(x_i)p(x_j)} \tag{5}$$

条件互信息定义为:

$$I(X_i,X_j \mid C) = \sum_{x_i,x_j,c} p(x_i,x_j,c)\cdot \lg\frac{p(x_i,x_j \mid c)}{p(x_i \mid c)p(x_j \mid c)} \tag{6}$$

式中,X_i、X_j表示 TM 影像的两个波段;$p(x_i)$表示 X_i的熵;$p(x_i,x_j)$表示 X_i、X_j的联合熵,它们都可以从影像(或样本数据)中采用统计的方法计算得到。根据式(5)和 § 2.1 节给出的样本数据,计算 TM 影像 6 个波段之间的互信息如表 1所示。

表 1　TM 波段之间的互信息

Tab. 1　Mutual Information Between TM Bands

互信息	波段 1	波段 2	波段 3	波段 4	波段 5	波段 6
波段 1	5.032 8	1.781	1.667 5	0.570 56	0.873 17	1.18 4
波段 2		4.501 8	2.258 6	0.697 24	1.134	1.445 7
波段 3			5.221 9	0.775 26	1.259 8	1.679 8
波段 4				5.503 6	1.266 2	1.029 2
波段 5					6.191 3	2.081 6
波段 7						5.696 8

表 1 中仅列出了上三角矩阵的值,其中主对角线的值是每个波段和自身的互信息。值越大,表示互信息越大,相关性越强;反之,独立性越强。

由信息论中的定义可知,两个随机变量仅当它们之间的互信息为零时相互独立。可见,NBC 特征节点间的条件独立是一种假设,其目的是为了简化联合概率的计算,但这种假设是不现实的。有两种改进 NBC 的方法,即特征选择(去掉相关性较强的特征)和放松条件独立性假设[1],后一种方法需要进行结构学习才能得到。从数据中学习得到结构是 NP 难题(non-deterministic polynomial problem),但如果对节点间的连接关系进行不同程度的限定,即可简化这个学习过程。如果将特征节点间的连接关系限定为树结构,即得到树增强型朴素贝叶斯网络(TAN);如果将特征节点间的连接关系限定为图结构,即得到贝叶斯增强型朴素贝叶斯网络(BAN)。

2.1　实验数据

以武汉市区 2005 年的 TM 影像 6 个波段(1、2、3、4、5、7 波段)为实验数据,选取 6 类样本:长江(412)、湖泊(405)、居民地(442)、林地(370)、裸地(326)、农用地(441),样本总数为 2 340。

由于遥感数据量大(256 灰度级被认为是连续变量),为提高学习效率,所有特征(波段的灰度值)要进行离散化处理。本文采用基于信息熵的离散化算法,该方法通过对离散特征的概括归纳

减少信息损失[9]。

实验中,采用5叠交叉检验方法得出分类器对样本的平均分类精度,并将这个精度作为分类器性能评价的标准之一。

2.2 基于互信息的结构学习

针对NBC简单条件独立性假设存在的问题,下面以TM影像波段间的互信息为度量,采用不同的结构学习算法分别得到上述三种分类器。

2.2.1 SNBC结构学习

采用后向选择算法获得SNBC的最优特征子集,具体过程如下:① 以全部波段作为初始特征集F_0,并获得其分类精度ρ;② 对波段间的互信息由大到小排序;③ 对互信息最大的两个波段中熵较小的一个波段X_i,去掉它与类节点的连接(但保留该节点),得到其分类精度ρ_1;④ 如果$\rho_1 \geqslant \rho$,从初始特征集中去掉X_i,否则恢复它与类节点的连接;⑤ 重复步骤③和④,直到$\rho_1 < \rho$,得到最优特征子集F_1,并记录当前的互信息值I_0。

通过这种方式确定一个合理的互信息阈值I_0,并认为互信息小于这个阈值的波段之间是相互独立的,对所有互信息大于这个阈值的两个波段,去掉熵较小的一个,即获得SNBC的最优特征子集F_1,进而得到其网络结构(见图3)。

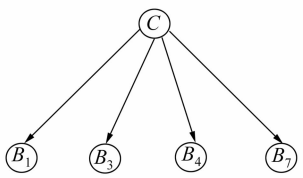


图3 SNBC的网络结构

Fig. 3 Network Structure of SNBC

2.2.2 TAN结构学习

TAN允许特征节点之间存在一定的相关关系,但将这种相关关系限制为一种树型关系。Chow and Liu提出用最大权重跨度树算法(maximal weighted spanning tree, MWST)来构造树[2],其算法原则是:在特征节点的完全图(节点两两之间都有连接)的基础上选择其连接的子集构成树结构,使得这些连接所对应的权重和最大,这里的权重用节点之间的互信息来表达。其学习步骤如下:① 根据式(6)计算节点对的条件互信息$I(X_i, X_j \mid C)(i \neq j)$;② 建立一个以所有节点为顶点的完全无向图,并为每条连接边赋权值,权值就是节点间的互信息,如节点X_i和X_j的连接边赋权值为$I(X_i, X_j \mid C)$;③ 在上述完全图的基础上用MWST算法[4]构造一棵最大权重跨度树(这个树仍是一个无向图,是步骤②中完全无向图的一个子集);④ 选择任意一个节点作为根节点,其他节点作为子节点,添加根节点到子节点间的连接,连接方向由根节点指向其子节点,将无向图转化成有向图。

用MWST算法学得的特征节点的树结构如图4所示。接下来,再将树中所有的节点作为类节点的子节点,添加类节点到特征之间(MWST算法生成的有向图中的所有节点均作为特征节点)的有向连接,组成完整的TAN结构。

2.2.3 BAN结构学习

BAN允许特征节点之间存在任意的连接关系,节点之间成为一种图结构。使用Cheng提出的条件独立性测试方法来构造贝叶斯网络[10]。条件独立性测试是根据节点间的互信息来判断两个节点是否条件独立,如果两节点间的互信息$I(X_i, X_j)$小于某个较小的阈值ε,就认为X_i和X_j是边缘独立的;如果$I(X_i, X_j \mid C) < \varepsilon$,就认为$X_i$和$X_j$是条件独立的[7]。一般取节点对互信息中一个较小的值(也可以取其中最小的值)作为阈值ε,以使得节点间的依赖关系被充分挖掘出来。

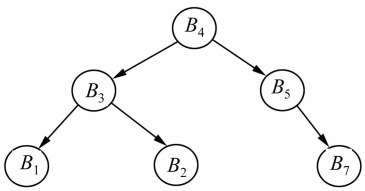


图4 TAN分类器特征节点的树结构

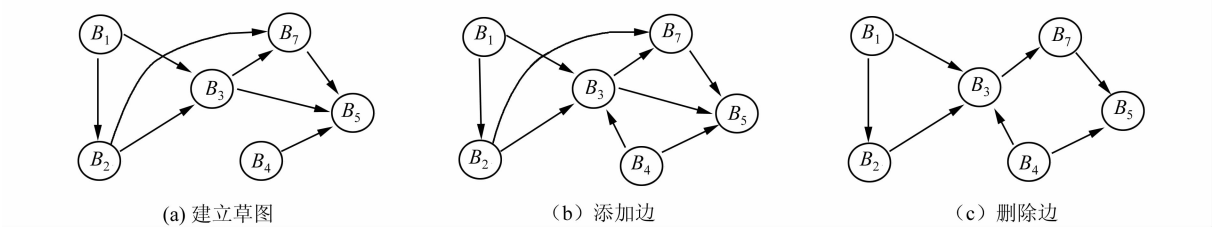
Fig. 4 Tree Structure of TAN Classifier's Feature Nodes

利用条件独立性测试方法来构造BAN的过程分为三步,即建立草图、添加边、删除边[7]。首先根据式(6)计算每个节点对间的条件互信息,在互信息较大(大于ε)的节点对间添加连接,画出网络结构草图;其次,对每个节点对进行条件独立性分析,如果不满足条件独立,增加连接边,然后再对每个节点对进行条件独立性分析,满足条件独立的节点对,去掉连接边;最后对所有的连接边确定其方向。

学习得到的BAN的特征节点的图结构如图5(c)所示。接下来再将图中的所有节点作为类节点的子节点,添加类节点到特征之间的有向连接,组成完整的BAN结构。

2.3 分类器性能及精度分析

为验证本文构建的分类器在遥感数据分类中的有效性,将NBC及其三种改进模型(SNBC、



(a) 建立草图　　(b) 添加边　　(c) 删除边

图 5　条件独立性测试方法构造 BAN 的三个步骤及 BAN 分类器特征节点的图结构

Fig. 5　Three Steps in Constructing BAN Utilizing Conditional Independence Test and Graph Structure of BAN Classifier's Feature Nodes

TAN、BAN)用于遥感数据的分类,并从分类器的计算复杂度和分类精度两方面进行评价。表 2 是 4 种分类器的参数及计算复杂度分析,连接边数指特征节点间有向边的个数。连接边数决定着参数个数,进而决定着模型的计算复杂度。计算复杂度的计算公式为 $\sum_{i=1}^{n} 2^{\pi(X_i)}+1$,其中,$n$ 为节点数;$\pi(X_i)$为节点 i 的父节点个数。

表 2　分类器参数及复杂度

Tab. 2　Parameters and Complexity of Classifiers

分类器	特征数	连接边数	参数个数	计算复杂度
NBC	6	0	13	—
SNBC	4	0	9	—
TAN	6	5	23	$O(N^2\lg N)$
BAN	6	7	31	$O(N^2)$

表3给出了4种分类器对样本(§2.1给出的实验数据)的分类精度。可以看出,SNBC 比 NBC 的精度提高了约 1%,这说明特征选择一定程度上消除了特征间的相关性。从计算代价(用参数个数和计算复杂度来衡量)来分析,SNBC 是最优的,但其分类精度相比 NBC 提高不大,可见去掉特征的同时,也损失了部分信息。总体上来看,TAN 和 BAN 比 NBC 和 SNBC 的精度有大幅度提高,这是因为 NBC 和 SNBC 特征节点间的条件独立性假设忽略了特征间的相关关系,而 TAN 和 BAN 将这种关系揭示出来,更符合实际情况。同时,TAN 和 BAN 的参数个数大幅度增加,计算复杂度更大(见表 3),且 BAN 获得了最好的分类性能(精度:95.93%;Kappa 系数:0.951 0),比 TAN 的精度提高了近 2%。并且除了农用地的分类精度和 TAN 基本一致,长江和湖泊两类地物的精度都很高,且基本一致外,其他各类地物的分类精度相比TAN,都有不同程度的提高。不同于TAN将特征节点间的连接关系约束为树型结构,BAN 对节点间的连接不加限制,以图的结构将特征间的相关关系充分揭示出来,能够发现 TAN 所不能发现的更多节点间的联系,从而获得更高的分类精度。

表 3　分类器的样本分类精度/%

Tab. 3　Classifier's Classification Accuracy for Samples/%

分类器	长江	湖泊	居民地	林地	裸地	农用地	总体精度	Kappa 系数
NBC	96.5	99.5	74.21	92.00	92.0	79.0	88.86	0.867 5
SNBC	99.75	99.75	68.92	95.79	91.72	83.00	89.82	0.877 1
TAN	99.25	99.58	80.77	96.53	93.69	94.80	94.10	0.935 2
BAN	99.25	99.25	88.0	98.47	95.30	94.78	95.93	0.951 0

参　考　文　献

[1] 陶建斌,舒宁,沈照庆. Naive Bayesian Classifier 在遥感影像分类中的应用研究[J]. 遥感信息,2009(2):57-61

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

[3] Madden M G. A New Bayesian Network Structure for Classification Tasks[C]. The 13th Irish International Conference on Artificial Intelligence and Cognitive Science, Irish, 2002

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

[5] Solares C, Sanz A M. Different Bayesian Network Models in the Classification of Remote Sensing Images[C]. Intelligent Data Engineering and Automated Learning-IDEAL 2007, Birmingham, UK, 2007

[6] 张连文,郭海鹏. 贝叶斯网引论[M]. 北京:科学出版社,2006

[7] 虞欣,郑肇葆,叶志伟,等. 基于 Tree Augmented

Naive Bayes Classifiers 的影像纹理分类[J]. 武汉大学学报·信息科学版,2007,32(4):287-289

[8] Cheng J, Greiner R. Learning Bayesian Belief Network Classifiers: Algorithms and System[J]. Lecture Notes in Computer Science, 2001, 2 056: 141-151

[9] 梁静,张桂峰. 基于信息熵的遥感影像特征离散化方法[J]. 地理空间信息, 2006, 4(3): 9-11

[10] Cheng J, Greiner R, Kelly J, et al. Learning Bayesian Networks from Data: An Information-theory Based Approach[J]. Artificial Intelligence, 2002, 137: 43-90

[11] Yu Xin, Zheng Zhaobao, Li Linyi, et al. Texture Classification of Aerial Image Based on PCA-NBC [C]. International Society for Optical Engineering, Wuhan, China, 2005

第一作者简介:陶建斌,博士生,现主要从事遥感影像智能化解译的研究。

E-mail:taojb_whu@163.com

An Improvement of Naive Bayesian Network Classifier for Remote Sensing Images Based on Mutual Information

TAO Jianbin[1] *SHU Ning*[1] *SHEN Zhaoqing*[1]

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

Abstract: This paper proposes an improvement of Naive Bayesian classifier-selective Naive Baysian classifier together with two enhancement of Naive Bayesian classifier-tree Augmented Naïve Bayes and Bayes Augmented Naïve Bayes. We constructed these classifiers for remote sensing images based on the mutual information between bands, and compared their performance with the NBC.

Key words: mutual information; analysis of conditional independence; Bayesian network classifier; remote sensing images; classification

About the first author: TAO Jianbin, Ph.D candidate, majors in intelligentized interpretation of remote sensing images.
E-mail: taojb_whu@163.com

(上接第 204 页)

Space Resection of Line Scanner CCD Image Based on the Description of Quaternions

YAN Li[1] *NIE Qian*[1] *ZHAO Zhan*[1]

(1 School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

Abstract: The theory of quaternions is introduced into the field of photogrammetry. A new method which uses quaternions to describe the position and attitude of line scanner CCD image is presented. Firstly the quaternions are used to describe the rotation matrix in the algorithm. Then, the strict collinear equation is linearized, and at the same time the iteration by correcting characteristic value can effectively overcome the strong interrelationship among exterior elements. The numerical experiments were done. Results show the correctness and reliability of the method.

Key words: quaterion; CCD image; space resection

About the first author: YAN Li, professor, Ph.D, Ph.D supervisor, majors in remote sensing image processing and 3D laser imaging radar.
E-mail: lyan@sgg.whu.edu.cn