遥感数据的贝叶斯网络分类研究

戴 芹[*] 马建文[*] 李启青[*] 陈 雪[*,**] 冯 春[***]
[*](中国科学院遥感应用研究所遥感信息科学开放实验室 北京 100101)
[**](北京师范大学遥感与地理信息系统研究中心 北京 100875)
[***](中国地质大学国土资源与高新技术研究中心 北京 100083)

摘 要: 由于遥感成像过程的复杂性，遥感数据中包含了一定程度的不确定性因素。利用最大似然分类器处理遥感数据时分类精度受一定的影响，为了提高分类精度往往需要引入先验知识。贝叶斯网络是一个带有概率注释的有向无环图，可以动态地对先验概率密度修正，提高分类精度，也没有严格的数据正态分布前提要求，适合处理不完整复杂的数据。该文介绍了利用贝叶斯网络对遥感数据进行分类处理的算法和技术过程。分类结果表明：贝叶斯网络具有稳定的数学基础，是一种可供遥感信息处理领域利用的有效新方法。

关键词：遥感数据，贝叶斯网络，分类
中图分类号：TP751 **文献标识码**：A **文章编号**：1009-5896(2005)11-1782-04

The Study on Remote Sensing Data Classification Using Bayesian Network

Dai Qin[*] Ma Jian-wen[*] Li Qi-qing[*] Chen-Xue[*,**] Feng Chun[***]
[*](*Institute of Remote Sensing Applications, CAS, Beijing* 100101, *China*)
[**](*Department of Resources and Environment Science, Beijing Normal University,Beijing*100875,*China*)
[***](*Institute of Land Resources and High Techniques, China University of Geosciences, Beijing* 100083, *China*)

Abstract Because of the complexity in satellite remote sensing imaging system, some uncertainties or mixed spectrum information are contained in the data. By using maximal likelihood classification to process remote sensing data, the result accuracy of the classification is affected. In order to improve the accuracy of the classification, prior knowledge is needed to modify the probability. Bayesian network is composed of directed acyclic graph and probability chart; it can modify the prior probability density dynamically and improve the accuracy of classification. In this paper, a technical procedure is demonstrated that using Bayesian network to process the remote sensing data, the classification results prove that Bayesian network has solid mathematics base and can be a new effective methods for remote sensing data processing.

Key words Remote sensing data, Bayesian network, Classification

1 引言

贝叶斯估计与最大似然估计同样具有稳固的数学基础，实验结果表明[1]，贝叶斯估计与最大似然估计相比突出特点表现在以下3个方面：(1)贝叶斯估计通过使用全部$P(\theta|\Pi)$中的信息，比最大似然估计方法取得更准确的结果；(2)数据曲线的不对称反映了数据的某些原始特征，贝叶斯估计能够利用这些特点，而最大似然估计采取了均值似然函数$P(\Pi|\theta)$的近似忽略了这些特点；(3)由于贝叶斯估计过程需要计算多重积分增加了运算的复杂性，特别是在处理数据量大、数据内涵相对复杂的遥感数据的应用进展比较缓慢。为此，许多论文都将注意力集中在解决和改进贝叶斯估计的选择特征数据集和在保证置信度和分类精度的前提下探寻放宽分类的假定条件。在不断的求证中发现，建立采用一种有向无环拓扑结构网络表达贝叶斯估计的基本原理，将变量间复杂的关系表达为连接概率的图形，使复杂的运算和推理表示为一种自然的因果信息网。贝叶斯估计网络还具有同步多特征流动态特点，成为当代智能处理引擎研究和数据挖掘方法的热点[2,3]。最典型的示例就是Cheng Jie等人[4]从可理解和可操作的角度出发开发的贝叶斯网络推理系统，这个系统将复杂的运算过程简单化、实用化[4]。贝叶斯网络计算引擎已经被Intel

2004-06-01 收到, 2005-01-13 改回
国家 863 项目(2003AA135080-2)和国家自然科学基金(40371086)资助课题

公司作为未来处理器架构的核心技术开展了研究[3]。大量的应用贝叶斯理论和算法处理遥感数据方面的论文主要出现在90年代，由于贝叶斯算法实现复杂过程，文章多数介绍的是个例研究结果很难普及使用[5,8]。随着贝叶斯网络的不断完善和发展，贝叶斯网络处理软件的开发，用户界面友好，将复杂的处理过程屏蔽，贝叶斯网络普及应用的时机已经到来。

由于遥感成像过程的复杂性，遥感数据表达地物光谱的多解性以及波段之间的相关性等原因导致利用最大似然分类器处理遥感数据时，分类精度普遍受到一定的影响，为了提高分类精度往往需要引入的先验知识。能否在处理过程中减少算法对先验知识的依赖，我们先后探索了自组织神经网络方法和基于辐射传输数据的方法等[9,10]。贝叶斯网络可能成为一种有效处理遥感数据的方法[11]。在贝叶斯网络方法中待估计的参数$\boldsymbol{\theta}$ 看成是一个符合某些先验概率$P(\boldsymbol{\theta})$的随机变量，对样本进行观测的过程是把先验概率密度转化成后验概率密度$P(\boldsymbol{\theta}|\Pi)$，将后验概率作为分类准则，这样就可以利用样本的信息修正参数的初始估计值。利用网络图形来表达类别与数据间，数据与数据间的相互关系，语义清晰易于理解。能否利用贝叶斯估计网络的这些优良性质解决遥感数据处理中对先验知识的依赖，是当前探索遥感数据贝叶斯估计分类的主要目标。

2　贝叶斯网络计算的实现过程

贝叶斯网络计算的实现过程主要包括以下几个步骤[11]：

(1) 确定网络模型目标，确定与目标有关的特征变量；

(2) 建立一个表达有向无环的网络结构图，计算联合概率，公式有

$$p(x)=\prod_{i=1}^{n}p(x_i|x_1,x_2,\cdots,x_{i-1})$$
$$=p(x_1)p(x_2|x_1)p(x_3|x_1,x_2)\cdots p(x_n|x_1,x_2,\cdots,x_{n-1}) \quad (1)$$

对于每个变量 x_i，如果有某个子集 $\Pi_i\subseteq\{x_1,x_2,\cdots,x_{i-1}\}$ 使得 x_i 与 $\{x_1,x_2,\cdots,x_{i-1}\}\setminus\Pi_i$ 是条件独立的，即对任何 x，有：

$$p(x_i|x_1,x_2,\cdots,x_{i-1})=p(x_i|\Pi_i)\quad i=1,2,\cdots,n \quad (2)$$

由式(1)和式(2)可以得到式(3)：

$$p(x)=\prod_{i=1}^{n}p(x_i|\Pi_i) \quad (3)$$

变量集合$(\Pi,\cdots,\Pi_n)$ 对应于父结点$(\mathrm{pa}_1,\cdots,\mathrm{pa}_n)$, 故又可以写成：

$$p(x)=\prod_{i=1}^{n}p(x_i|\mathrm{pa}_i) \quad (4)$$

于是为了决定贝叶斯网络的结构，需要将变量 $x_1,x_2,\cdots,x_i$ 按照某种次序排序，决定满足式(2)的变量集；

(3) 设置局部概率分布 $p(x_i|\mathrm{pa}_i)$。在离散的情形，需要为每一个变量 x_i 的各个父结点的状态指定一个分布；

(4) 贝叶斯网络的学习。在给定贝叶斯网络的结构，利用给定的样本数据去学习网络的概率分布，更新网络变量原有的先验分布。假设变量组 $x=(x_1,x_2,\cdots,x_{n-1})$ 的联合概率分布可以编码在网络结构中：

$$p(x|\boldsymbol{\theta}_s,s^h)=\prod_{i=1}^{n}p(x_i|\mathrm{pa}_i,\boldsymbol{\theta}_i,s^h) \quad (5)$$

其中 $\boldsymbol{\theta}_i$ 是分布 $p(x_i|\ \mathrm{pa}_i,\boldsymbol{\theta}_i,s^h)$ 的参数向量，$\boldsymbol{\theta}_s$ 是参数组 $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\cdots,\boldsymbol{\theta}_n)$ 的向量，而 s^h 表示物理联合分布可以依照 S 被分解的假设。

(5) 利用验证数据对训练后的网络进行验证，程序中设定置信度为 95%。

3　实验与分析

为了实验贝叶斯网络的遥感数据分类方法，我们选择了2003 年 5 月 1 日北京南部地区的 6 个波段的 ETM+数据，波段 1，2，3，4，5，7，实验数据为 400×400 个像元。图 1 是 ETM+5，4，3 波段的合成影像图，5 波段为短波红外波段，4 波段为近红外波段，3 波段为可见光红波段，在影像中黑色为水体，大片的灰色为农田，较深灰色为林地，白色为荒地，浅灰色为裸露的地面，深灰色为建筑用地。

利用贝叶斯网络分类的技术路线选择与步骤，见图 2。

(1) 选择 ETM+ 6 个波段遥感数据，波段 1, 2, 3, 4,5, 7；

(2) 选取训练数据集和验证数据集(特征向量)，见表 1, 选择样本时通过对照假彩色合成图与实地调查的指导下进行；

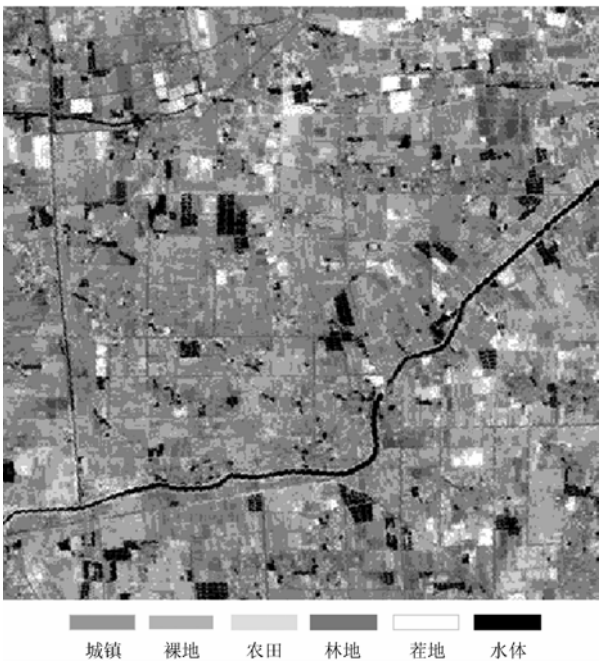

图 1　实验区 5、4、3 波段(RGB)合成图

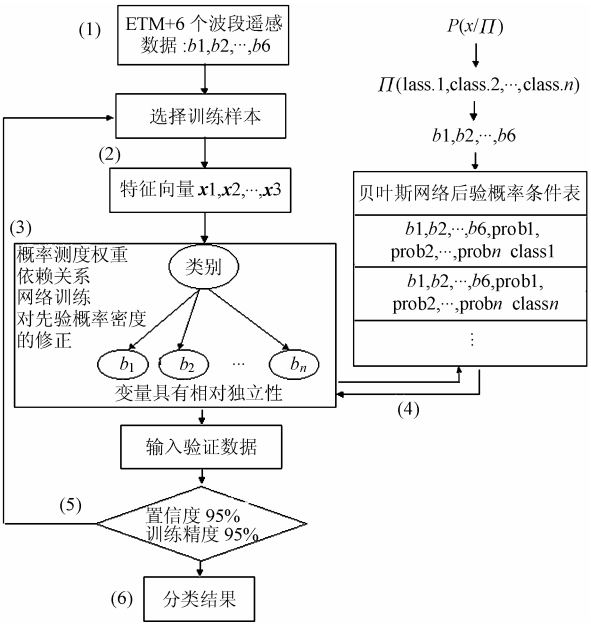

图 2 遥感数据贝叶斯网络分类流程图

表 1 训练数据集合与验证数据集合

类别号	土地覆盖	训练数据集合	验证数据集合
1	城镇用地	120	55
2	裸地	135	24
3	农田	80	23
4	林地	60	18
5	荏地	75	24
6	水体	60	16
合计		530	160

(3) 建立贝叶斯网络分类结构；

(4) 应用训练数据集依照建立好的网络结构，确立概率测度权重依赖关系，对数据进行训练，得出概率条件表，概率测度权重依赖关系，网络训练对先验概率密度的修正；

(5) 在置信度为 95%的条件下，应用验证数据集对网络分类精度进行评价，如果满足分类精度的要求，可以对新的数据进行分类；

(6) 网络分类结果见表 2，图 3。

表 2 是贝叶斯网络的分类结果的 6 种类别的混淆矩阵，表达了贝叶斯网络在置信度为 95%的条件下，由于本文侧重于方法实验，因此在选择区域地物类型简单、明确，在 160 个验证点的数据集中，总体精度达到 100%。

图 3 贝叶斯网络分类图，图中黑色为水体，深灰色为农田，较深灰色为林地，灰色为荏地，白色为建筑用地，浅灰色为裸地。

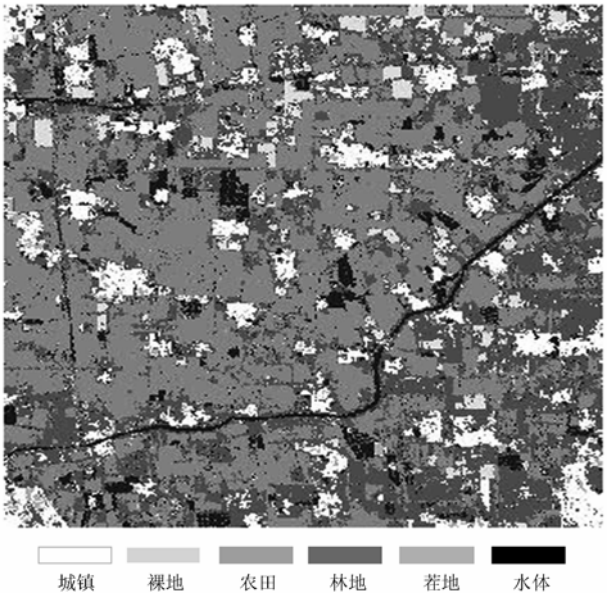

图 3 贝叶斯网络分类结果图，共 6 类。

4 结束语

贝叶斯网络软件的开发和友好的界面，为用户提供了方便，它不需要经过复杂的数学和推理过程就可以使用贝叶斯网络处理面临的实际问题。本文介绍了利用贝叶斯网络对 ETM+的 6 个多光谱波段数据的分类过程，实验中展示了贝叶斯网络采用图形的方式描述数据间的相互关系，它语义清晰、易于理解，为遥感数据分类提供了一种可选择途径；这个实现过程还没有能在贝叶斯网络实现过程(3)，(4)中充分体现数据网络节点(多特征)之间的连接关系和运算，这一运算的重要性将在利用多源空间数据约束网络训练，提高分类精度方面充分体现。

表 2 贝叶斯网络的分类结果的混淆矩阵

类别	城镇用地	裸地	农田	林地	荏地	水体	精度
城镇用地	55	0	0	0	0	0	0.83
裸地	0	24	0	0	0	0	0.93
农田	0	0	23	0	0	0	0.93
林地	0	0	0	18	0	0	0.94
荏地	0	0	0	0	24	0	0.93
水体	0	0	0	0	0	16	0.95

参 考 文 献

[1] Richard O D, Peter E H, David G S 著,. 李宏东,姚天翔等译, 模式分类. 北京：北京机械工业出版社, 2001: 16－81.

[2] 周颜军，王双成，王辉. 基于贝叶斯网络的分类器研究. 东北师大学报自然科学版, 2003, 35(2): 21－27.

[3] Bob Liang. 未来处理器架构进行应用驱动的研究报告,Intel Microprocessor Research 2002 Forum, Beijing China , October 29, 2002: 1－20.

[4] Cheng, J, Greiner R, Kelly J, Bell D A, Liu W. Learning Bayesian networks from data: an information-theory based approach. *The Artificial Intelligence Journal*, 2002, 137(1): 43－90.

[5] Hurn M A, MardiaKV. Bayesian fused classification of medicalImages. *IEEE Trans. on Geoscience and Remote Sensing*, 1999, 37: 1292－1305.

[6] 哈斯巴干, 马建文, 李启青, 韩秀珍, 刘志丽. 基于小波变换的 ASTER 数据的自组织特征映射神经网络分类研究, 中国科学(D 辑), 2003, 33(9): 896－902.

[7] Ma Jianwen, Guo Huadong, Wang Changlin, *et al*.. Extraction of polymetallic mineralization in formation from multi-spectral thematic mapper data using the Gram-Schmidt orthogonal projection (GSOP) method. *Int. J.Remote Sensing*, 2001, 22 (17): 3323－3337.

[8] 李启清，马建文，哈斯巴干等. 基于贝叶斯网络模型的遥感数据处理技术. 电子与信息学报, 2003, 25(10): 132－136.

[9] 范明, 孟小峰等. 数据挖掘概念与技术. 北京：机械工业出版社, 2001: 196－200.

[10] Cooper G F, Herskovtis E. A Bayesian method for the induction of probabilistic network from data. *Machine Learning*. 1992, 9(4): 309－347.

[11] 史忠植. 知识发现. 北京：清华大学出版社, 2002: 169－198.

戴　芹：女，1978 年生，博士生，研究方向为遥感信息处理与应用.

马建文：男，1953 年生，博士，博士生导师，创新研究员，遥感技术发展部主任，加拿大留学，中国科学院 1998－2001 百人计划，国际 IEEE 会员，中国遥感学会理事。INT.J Reviewer,主要从事遥感数据模型和信息处理，发表论文：SCI 6 篇、EI+CSCD 70 余篇,专著：3 部.

李启青：男，1977 年生，北京大学博士后，研究方向为遥感图像处理.

陈　雪：女，1977 年生，博士生，研究方向为遥感图像智能处理.

冯　春：男，1978 年生，博士生，研究方向为遥感图像处理.