基于贝叶斯网络模型的遥感图像数据处理技术 [1]

李启青 马建文 哈斯巴干 韩秀珍 刘志丽

(中国科学院遥感应用研究所技术部 北京 100101)

摘 要 贝叶斯网络是一种不确定性知识的推理和描述技术，针对遥感数据的复杂性和不确定性，该文提出了一种基于贝叶斯网络模型的遥感数据推理和描述技术. 文中利用 2002 年春季中 - 日亚洲沙尘暴项目的土地利用数据 (LU), 沙尘监测数据 (TSP), 卫星 AVHRR 时间序列 LST/Albedo 数据，采用贝叶斯网络模型进行了知识描述和信息推理预测实验，取得了较好的效果.

关键词 贝叶斯网络模型，知识描述，信息推理，遥感图像数据

中图号 TP751

1 前 言

遥感图像处理中经常要面对不充分的数据和信息，应该如何对这些数据和信息进行组织描述和应用推理是数据挖掘和信息提取领域的重要课题. 作为一种信息推理技术，贝叶斯网络模型 (Bayesian Network Model, BNM) 可以很好地解决这个问题. BNM 也被称为条件概率网络模型 (Conditional Probability Network Model ,CPNM) 或者原因概率网络模型 (Reason Probability Network Model ,RPNM) . BNM 是一个图形模型，该模型提供了描述变量之间概率关系的一种方式；同时，在条件独立假设和局部概率分布的前提下，描述了变量之间的联合概率分布 [1] . 这种模型的好处在于使用图形结构描述输入属性之间相互关联的方式. 这样， BNM 可以被用于非同源遥感数据，合并不同的先验知识和数据信息.

BNM 是近年来研究贝叶斯统计方法的新进展，是一种新的推理技术，是图论和概率论结合所产生的一种信息描述方法 [2] . BNM 主要用于复杂多因果关系的分析，它使用概率理论来处理在描述不同知识成分之间的条件相关而产生的不确定性 [3] .

BNM 说明了联合条件分布，允许在变量的子集之间定义类条件独立性 [4,5] . 它提供了一种因果关系图形，可以在其上学习并根据学习结果进行分类推理设置预测. 它克服了朴素贝叶斯方法无法定义变量之间的依赖关系的弱点 [6] . BNM 既是一种信息描述方式，又是一种推理技术 [7] . BNM 同时也是学习数据分类的一种基本技术，是人工智能技术的重要领域 [8] . 微软是 BNM 研究应用最积极的探索者之一，在使 BNM 能自动从新知识中学习或更新的技术的研究和使用 BNM 技术改进从大型数据库中查找相关信息片段的人工智能技术的研究中具有独特之处 [8] . 遥感图像数据所具有的不确定性使得其在应用时表现出较强的复杂度 [9], BNM 处理技术在用于解决这种复杂性问题的描述，表达和信息推理方面具有独特的优势.

本文首先简单介绍了贝叶斯定理和贝叶斯网络，然后介绍了根据样本数据和先验信息建立贝叶斯信念网络的过程. 针对遥感图像数据的不确定性，提出并详细介绍了遥感数据处理领域中 BNM 的应用方法. 通过对 AVHRR 数据以及相关土地利用类型，沙尘干量数据的处理实验，并实际进行了沙源区起沙程度预测应用分析实验. 本方法对在大型遥感图像数据库上运用贝叶斯信念网络进行数据分析和信息推理有重要的现实意义.

2 贝叶斯网络模型

传统的数据统计技术完全立足于“单纯，死板”的数据信息，而以贝叶斯定理为理论基础的

[1] 2002-09-23 收到， 2002-12-23 改回
国家科技攻关项目 (编号： 2002BA904B07) 和 973 项目 (编号： G2000077904-2) 资助

数据统计技术有机地将数据信息与真实世界的信息 (先验信息) 联系在了一起 [10] 。

除了提供一种计算后验概率的方法，贝叶斯定理的优势还在于能够帮助人们建立起分析复杂真实世界的模型——贝叶斯网络。这种更容易令人理解和把握的图形模型对复杂和不确定的信息具有很强的处理能力。它由两部分组成，即有向无环图 (DAG) 和条件概率表 (CPT)，如 (1) 式所示：

$$B =< G,\theta > \tag{1}$$

给定一个随机变量集 $\chi = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n\}$，其中 X_i 是一个 m 维向量。贝叶斯信念网络说明了变量集合 χ 上的一个联合条件概率分布。

2.1 有向无环图 (Directed Acyclic Graphs, DAG)

(1) 式中 G 表示一个有向无环图，其顶点对应于有限集 χ 中的随机变量 $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ 。其弧代表函数依赖关系。如果有一条弧由变量 Y 到 X，则 Y 是 X 的双亲或者直接前驱，而 X 则是 Y 的后继。一旦给定其双亲，图中的每个变量独立于图中该节点的非后继。 $\boldsymbol{X}_i$ 的所有双亲变量用集合 $P_a(\boldsymbol{X}_i)$ 表示。

2.2 条件概率表 (Condition Probability Table, CPT)

(1) 式中 θ 代表量化网络的一组参数。对于每一个 $\boldsymbol{X}_i$, $P_a(\boldsymbol{X}_i)$ 的取值 x_i 存在如下一个参数： $\theta_{x_i|P_a(\boldsymbol{X}_i)} = P(x_i|P_a(\boldsymbol{X}_i))$，它指明了在给定 $P_a(\boldsymbol{X}_i)$ 发生的情况下 x_i 事件发生的条件概率。

因此，该贝叶斯信念网络给定的变量集合 χ 上的联合条件概率分布为

$$P_B(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n) = \prod_{i=1}^{n} P_B(\boldsymbol{X}_i|P_a(\boldsymbol{X}_i)) \tag{2}$$

3 遥感数据信息的贝叶斯网络描述

遥感数据信息具有的不确定性使得描述它的贝叶斯网络也具有特殊性，这种特殊性表现在数据组织方式以及贝叶斯网络结构及参数的确定过程中，并且针对不同的具体问题应该有不同的贝叶斯网络表达形式。

3.1 信念

信念是指根据自己生活经历的积累对该事件发生的可能性所给出的确认程度，在贝叶斯统计中信念也称为主观概率或者先验信息。主观概率的确定要求当事人对所考察的事件有较透彻的了解和丰富的经验，在这个基础上确定的主观概率就能符合实际 [10] 。在我们的贝叶斯网络模型中关于沙源区有如下先验信息：

近几年每年三，四月份经常发生大的沙尘暴，其波及面比较广。根据前人的研究成果起沙区主要分布在毛乌素沙漠，科尔沁沙漠，塔里木沙漠等地。

3.2 贝叶斯网络结构及其确定

贝叶斯网络结构主要指一个有向无环图模型，图中的结点与相关变量一一对应，将结点联结起来的弧对应这些变量之间的联合概率分布。在这个模型中，结点表示的随机变量代表世界上的事件或事物，它们之间的影响程度通过一个数字编码的概率值来表示。一般而言，确定贝叶斯网络模型一般采用贝叶斯网络修正算法或者更新算法 [11] 。

根据沙尘源区预测的需要通过如下 3 步建立沙尘过程的贝叶斯网络模型。

(1) 确定为建立模型有关的变量及其解释。

lst 为前一天的地表温度； R_1 为前一天的地表反照度； LU 为当年的土地利用类型； tsp1 为沙源区的沙尘干量； tsp2 为其它地区的沙尘干量； S 为预测的起沙程度。

(2) 通过分析变量之间的条件依赖关系构建贝叶斯网络，即建立有向无环图 (图 1) 。

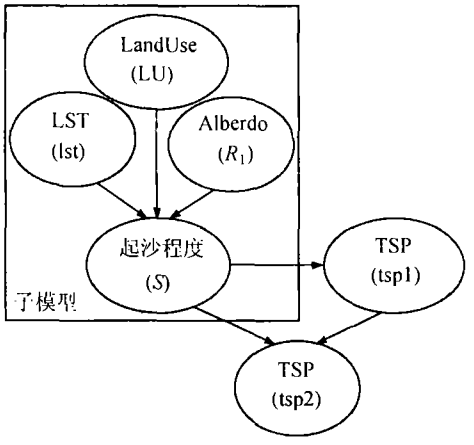

图 1　贝叶斯网络模型的有向无环图

根据概率乘法公式有

$$p(x)=\prod_{i=1}^{n}p(x_i|x_1,x_2,\cdots,x_{i-1}) \tag{3}$$

用 P_{ai} 表示变量 $\boldsymbol{X}_i$ 的父结点集，则

$$p(x)=\prod_{i=1}^{n}p(x_i|P_{ai}) \tag{4}$$

(3) 分析确定各个变量 (结点) 的条件概率表，即指派局部概率分布。在离散的情形，需要为每一个变量 $\boldsymbol{X}_i$ 的父结点集的各个状态指派一个分布 [10]。

比如在我们所建立的贝叶斯网络模型中，给定变量 LU, lst, R_1 的先验概率，服从高斯分布，变量 S 的父结点 R_1, lst, LU 的概率分布分别由其条件概率表表示。

3.3 条件概率表和参数学习

经过调查可用的遥感数据有 3 月 15 日，3 月 16 日，3 月 19 日，3 月 20 日，3 月 21 日，3 月 24 日。由表 1, 图 2 可知毛乌素沙区 3 月 17 日，3 月 25 日 TSP 值高于平均值，可判定为强起沙日。3 月 16 日，3 月 20 日，3 月 21 日，3 月 22 日的 TSP 值远低于平均值 280.53, 可判定为不起沙日。同样，3 月 19 日科尔沁沙地 TSP 为 1869.69, 属于强沙尘天气，并且影响了我国东部大部分地区。

表 1　IEECAS 仪器所测毛乌素沙区的 TSP 数据表 (TSP 单位：$\mu g/m^3$)

月－日	3-15	3-16	**3-17**	3-19	3-20	3-21	3-22	3-23	**3-25**	3-26	平均值
TSP	164.05	133.42	**1210.19**	209.86	31.50	119.97	135.67	160.20	**300.73**	223.24	280.53

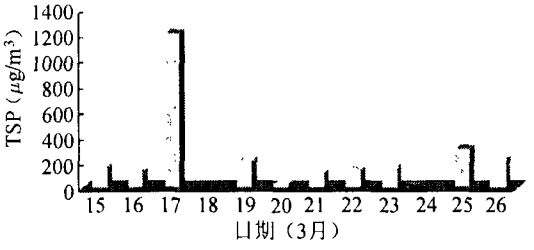

图 2　2002 年毛乌素沙区 3 月 15 日至 3 月 26 日的 TSP 数据

如表 2 所示，分别根据 3 月 17 日的前日 (3 月 16 日), 3 月 25 日的前日 (3 月 24 日) 和 3 月 16 的前日 (3 月 15 日), 3 月 20 日的前日 (3 月 19 日), 3 月 21 日的前日 (3 月 20 日), 3 月 22 日的前日 (3 月 21 日) 的 lst 和 R_1 的统计数据 (表 3) 可以得出变量 S 的统计表，并进而得到其条件概率表。

需要注意的是，根据需要，每天的统计数据需要选取无污染像元 5 个点左右，本实验笔者选取了 5 个点 (去除了一些污染像元)。

表 2 沙尘源区训练点选取的日期统计 (2002 年)

	数据统计日	沙尘预测日	验证
起沙	3 月 16 日	3 月 17 日	
	3 月 19 日 (科尔沁)	3 月 20 日 (科尔沁)	3 月 20 日 (科尔沁)
	3 月 24 日	3 月 25 日	
不起沙	3 月 15 日	3 月 16 日	
	3 月 19 日 (毛乌素)	3 月 20 日 (毛乌素)	3 月 20 日 (毛乌素)
	3 月 20 日	3 月 21 日	
	3 月 21 日	3 月 22 日	

注意： 3 月 18 日和 3 月 24 日无 TSP 数据

4 沙尘源区起沙指数的贝叶斯网络推断实验

根据所建立的贝叶斯网络模型 (图 3), 其变量的子集所组成的子模型 (图 1) 可以用来进行沙源区起沙程度的分类推断。模型中采用的地表温度，反照度图像数据使用多种方法进行反演，但反演方法并未采用遗传算法进行优化 [9]。

起沙程度变量 S 由 LU, lst, R_1 确定， lst 及 R_1 的统计结果见表 3, 根据这个结果进行的训练模型对某一天 (比如 3 月 20 日) 沙源区的起沙程度进行分析推断 (使用 3 月 19 日的数据), 得到的结果图像如图 4。

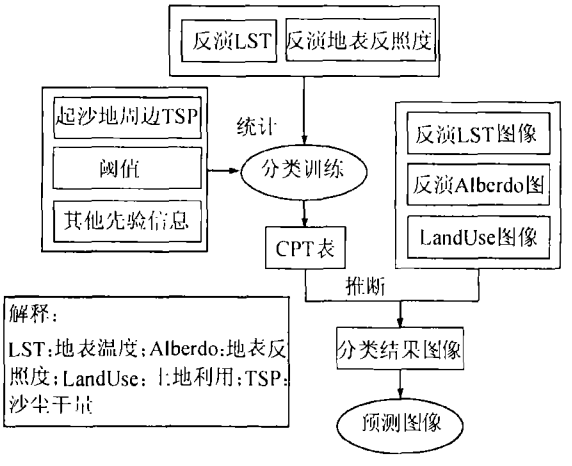

图 3 遥感图像数据的贝叶斯网络推断流程图

表 3 变量 R_1 和 lst 的统计数据表

S	$R_1 \le 0.1$	$R_1 \le 0.1$	$0.1 < R_1 \le 0.15$	$0.1 < R_1 \le 0.15$	$R_1 > 0.15$	$R_1 > 0.15$
	lst≤300	lst>300	lst≤300	lst>300	lst≤ 300	lst>300
有沙尘				5		5
无沙尘	2	4	1	1	6	4

注意：表中 R_1 表示地表反照度， lst 表示地表温度；表中数据表示点数。

沙源区预测图像与土地覆盖类型图叠合形成的合成图像如图 4 所示。其中 A 区域为毛乌素沙漠，B 区域为科尔沁沙漠，由此图像可以看出 3 月 20 日大规模起沙区域主要分布在科尔沁沙区，由此可以推断 3 月 20 日的沙源主要为科尔沁沙漠。

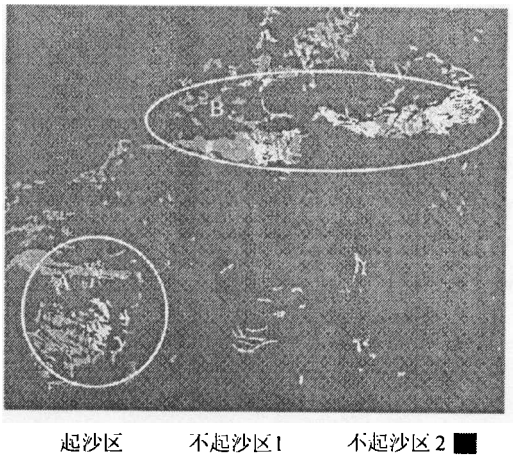

图 4　2002 年 3 月 20 日沙源区起沙情况预测合成图像 (毛乌素沙漠 A 和科尔沁沙漠 B)

5 结论与讨论

本文提出的贝叶斯网络模型实际上是一种多源遥感数据的挖掘方法，不同信息和数据之间的结合所具有的复杂性需要贝叶斯网络这样的智能处理工具。作为一种新的概率推理技术，贝叶斯网络模型结合了图论表达清晰和贝叶斯统计学的现实合理性，可以有效地解决遥感数据的处理问题。

实际工作中，对数据的组织统计仍需要继续探讨，比如选取训练统计点的个数可以增多。实际的预测工作可以选取更长的时间链条，比如通过昨天加前天甚至更前天的数据来统计组合，然后对今天甚至明天的沙尘情况进行预测。也可以预测不仅仅是沙尘源区的情况，当然这需要更多的变量。这既体现了遥感数据对时间的敏感性，也体现了 BNM 推理的因果影响延时效应。需要提及的是，本文所用的 BNM 有一个有向无环图 (DAG) 假定并且没有引入隐含变量，也没有做进一步的优化处理，这是基于降低推理复杂度的考虑 [12]。

对 2002 年的沙尘暴进行的详细的描述和推理试验证明本文所提供的方法具有较强的可靠性和实用性，进一步的验证将会有另文阐述。

参　考　文　献

[1] D. Heckerman, Bayesian networks for data mining[J], Data Mining and Knowledge Discovery, 1997, 1(1), 79–119 .

[2] 胡玉胜，涂序彦，等，基于贝叶斯网络的不确定性知识的推理方法，计算机集成制造系统，2001, 7(12), 65–68.

[3] 慕春棣，戴剑彬，等，用于数据挖掘的贝叶斯网络，软件学报，2000, 11(5), 660–666.

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

[5] Liu Zhiqiang, Causation, Bayesian networks and cognitive maps, 自动化学报, 2001, 27(4), 552–566.

[6] 林士敏，田凤占，陆玉昌，贝叶斯网络的建造及其在数据采掘中的应用，清华大学学报 (自然科学版), 2001, 41(1), 49–52.

[7] 王君圣，李敏强，基于数据库信息构建贝叶斯网络的 GA 方法，系统工程与电子技术，2000, 22(9), 54–57.

[8] 王军，周伟达，贝叶斯网络的研究与进展，电子科技，1999, 15, 6-7.
[9] 马建文，等，遥感数据模型与处理方法，北京，中国科学技术出版社，2001, 第 5 章第 2 节.
[10] 茆诗松，编著，贝叶斯统计，北京，中国统计出版社，1999, 第 1 章.
[11] 庄家礼，陈良富，等，用遗传算法反演连续植被的组分温度，遥感学报，2001, 5(1), 1-8.
[12] 钟清流，BN 结构和参数学习算法改进，微型电脑应用，2001, 17(5), 8-10.

A PROCESSING METHOD FOR REMOTE SENSING IMAGERY DATA BASED ON BAYESIAN NETWORK MODEL

Li Qiqing Ma Jianwen Hasi Bagan Han Xiuzhen Liu Zhili

(*Institute of Remote Sensing Application, Chinese Academy of Sciences, Beijing* 100101, *China*)

Abstract Bayesian network is a new inference and express method of uncertain knowledge. It is proposed an inference and express technique for remote sensing imagery data which has complexity and uncertainty based on Bayesian Network Model(BNM). In the paper, the LU data, TSP and LST/Albedo data of AVHRR time-sequence imagery which get from the project of China-Japan Asian dust storm in 2002 are used to analyze the dust storm and at the same time BNM is used to describe the knowledge and information inference. The satisfied results are given in the paper with the method.

Key words Bayesian Network Model(BNM), Knowledge description, Information inference, Remote sensing imagery data

李启青： 1977 年生，男，博士生，主要研究领域为图像处理，遗传算法和贝叶斯网络.
马建文： 1953 年生，男，研究员，博士生导师，研究兴趣为遥感应用模型与方法研究.
哈斯巴干： 1967 年生，男，博士生，主要研究领域为遥感数据模型与处理算法.
韩秀珍： 1974 年生，女，博士生，主要研究领域为遥感图像处理.
刘志丽： 1974 年生，女，博士生，主要研究领域为地图学与地理信息系统.