第32卷 第16期 计 算 机 工 程 2006年8月
Vol.32 № 16 Computer Engineering August 2006
·开发研究与设计技术· 文章编号：1000—3428(2006)16—0271—03 文献标识码：A 中图分类号：TP391

面向遥感图像 BNs 分类的预处理技术及算法实现

李启青，程承旗，郭仕德，何华伟
（北京大学遥感与地理信息系统研究所，北京 100871）

摘 要：遥感图像 BNs 分类分为预处理、BNs 模型构建和分类 3 个前后联系的过程。其中对预处理技术是后面两个步骤的基础，其算法实现过程对分类结果的影响很大，预处理的目标是高效、准确地提取图像分类所需要的重要特征，剔除干扰因素。该文介绍了一种简单数据预处理技术和其算法实现过程。针对遥感数据的特点和 BNs 方法的需要，将该预处理过程分成波谱空间分割和关系信息计算两个部分，分别介绍了两部分的原理并给出了实现的算法。对于遥感数据分类预处理技术的研究和实现具有很强的借鉴作用。
关键词：遥感图像；分类；贝叶斯网络；预处理；实现

A Preprocessing Technique for Remote Sensing Bayesian Networks Classification and Its Algorithm Implementation

LI Qiqing, CHENG Chengqi, GUO Shide, HE Huawei
（Institute of RS & GIS of Peking University, Beijing 100871）

【Abstract】 BNs classifications for remote sensing image is divided into three courses include preprocessing, the BNs model construction and image classification. Preprocessing is the foundation of the other two steps. The preprocessing step has large effect on the whole classification results. The goal of preprocessing is to extract the information high-efficiency and exactly and to eliminate disturbance from image features. This paper introduces the process about data preprocessing technique and the algorithm implementation. Because of the characteristic of remote sensing data and BNs method, the preprocessing is divided into two parts, one is spectrum space segmentation, and the other is mutual information computation. The principles and the algorithms of two parts are introduced. It is very important for the development of preprocessing technique.
【Key words】 Remote sensing image; Classification; Bayesian networks; Preprocessing; Implementation

遥感数据具有不确定性，这种数据的不确定主要是由数据获取的过程引入的。在数据获取以后，要获得关于地物的信息就需要去除上述不确定性。为了获得地物的类别信息，通过遥感图像数据进行自动化分类是目前有效并且合理的手段。贝叶斯网络作为一种分类手段用于遥感图像具有独特的优势，它可以采用概率的方式描述确定的图像数据中涵盖的不确定性，还具有表达大量属性之间条件独立关系的图形化方式。要合理地描述遥感数据中的不确定性，必须选择一种有效的数据预处理技术，并高效地实现这种技术。

数据分类预处理的目的不但在于提高数据的质量从而提高分类结果的质量，而且可以使得分类过程更加有效、更加容易，降低实际分类器训练、分类等过程的时间。针对遥感数据的 BNs 分类器的建造过程而言，虽然从遥感图像所获得的训练数据本身已经是离散数据。但是其粒度太低，不能够满足建造过程的要求，因此针对 BNs 分类所做的数据预处理就要提高数据的粒度。

在提高训练数据的粒度以后，要从中获得BNs遥感图像分类模型，可以使用波段之间的关系信息[1, 2]。这就需要一个BNs方法所特有的预处理过程——获取波段之间的关系信息(mutual information)。关系信息的度量方法，下文将会给出相应的实现过程。

1 波谱空间分割

因为遥感数据的多波谱(也有称为多光谱)特性，提高数据粒度的过程对于遥感数据可以称为波谱空间分割，也可以称为离散归一化。构建合理的BNs模型实际上是一个从训练数据中提取模型信息的过程。训练数据中，定量的数据对要表达的地物类别信息而言，是多对一的关系，相对来说，描述是不精确的。要将这些描述作为数据库，从这样的数据库中提取有用信息，从有用信息中发现知识，从知识中推理决策规则，就是整个模型构建过程。将光谱数据进行离散归一化则是整个过程的基础，因为作为理论方法分析的决策基础(BNs模型)是有限维的离散化数据表。离散化的方法应该满足的要求有：(1)属性离散归一化后的空间维数尽量小，也就是每一离散归一化后的属性值的种类尽量少；(2)属性值被离散归一化后的信息丢失尽量少；(3)效率要尽量高。对于不同的数据如定性说明型数据值、定性说明有层次分别型的数据值、连续值的数据值离散归一化的具体做法都有不同[3]。

离散归一化的方法主要分为两类：(1)对每一个属性的属性值进行划分的局部离散方法；(2)同时考虑全部条件属性的属性值进行划分的全局离散方法。前者效率较高，后者由于要同时考虑全部条件属性的属性值，效率较低。根据上述要求，我们选择了效率比较高同时较简单的第(1)类方法。局部离散归一化方法采用如下定义：

对于一个波段 a，设波段数据值的域为 π_a，离散归一化就是产生一个对波段数据值的域的划分：

基金项目：中国博士后科学基金资助项目(2005037276)；国防“863”计划基金资助项目
作者简介：李启青(1977—)，男，博士后，主研方向：遥感图像处理；程承旗，教授、博导；郭仕德，高工；何华伟，硕士
收稿日期：2005-09-03 **E-mail：**liqiqing@postdoctor.org.cn

$$\pi_a = \{[d_0, d_1], [d_1, d_2], \cdots, [d_{k-1}, d_k]\}$$

其中，$d_0 = a_{\min}$，$d_k = a_{\max}$，$d_{i-1} < d_i, i = 1, \cdots, k$，$i$就是离散归一化代表值，$k$是离散归一化的级。最简单的方法是等间隔划分法：将 $(a_{\max} - a_{\min}) / k$ 作为划分间隔，从而离散归一化该波段的所有“连续”数据值。第 2 种简单的局部离散归一化方法是等频率间隔划分法，即从 $a_{\min}$ 开始，每次取相同数目的属性值样本作为一个间隔，若该属性的属性值总数目为m，离散为k级，则每一个间隔中的样本数目为 m / k。除此之外划分间隔断点的方法还有利用分类熵、最小距离等准则。还有一种进行波段数据值离散归一化处理的方法称为全局聚类分析法[3]。等宽度动态分割波谱空间的步骤如图 1。

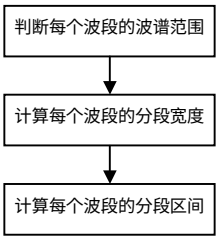

图 1 等宽度动态分割波谱空间的 3 个步骤

2 数据归类统计

为了最终判断每个分割空间所拥有的训练数据点数，首先需要判断每个数据是属于哪个分割区间并进行标号，为此，本文给出了一种实现。

2.1 原理

对每个波段上的最大值与最小值之间采用等宽度方法来划分属性值区间，并对每个像素点的每个波段上的值分别判断其所属的区间范围。

具体方法是使用一个数组 N，按照波段的顺序和每个波段属性值从小到大的顺序依次存储了所有的分段区间，N 的一个数据的格式为“区间内最小值<；<区间内最大的值”，使用数组 N2 来记录每个波段的属性值区间的个数。

对于某条记录的某个波段上的值 P，可以通过以下公式来计算它在该波段的第几个属性值区间内。

$$\mathrm{nRange}_{Bi} = \frac{(\mathrm{P} - \mathrm{Bi波段上最小值})}{\mathrm{Bi波段的属性值区间宽度}} (\text{i 表示第几个波段}) \quad (1)$$

当求出 nRange_{Bi} 之后，可以通过下面的公式计算得到的 n，n 表示对应到数组 N 的第几个记录，从而就可以得出 P 在第 i 个波段上所属于的属性值区间了。

$$n = \mathrm{nRange}_{Bi} + N2_1 + N2_2 + \ldots + N2_{i-1} \quad (2)$$

2.2 算法步骤

(1)通过公式 m=(所有字段数-6)/3 可以求出波段数 m 的值。

(2)从用户选择的每个波段的属性值区间的个数存入到数组 N2 中。

(3)通过 SQL 语句检索出每个波段的最大值和最小值，并存入到数组 N_{max} 和 N_{min} 中。

(4)根据 N_{max}，N_{min} 以及 N2 可以计算出每个波段的每个属性值区间，并按照波段排列和属性值从小到大排列，并将结果存入数组 N 中。

(5)把数据库指针移动到第 1 条记录

(6)从数据库中取第 j 条记录，j∈(第三者记录,...,最后一条记录)。

(7)从 i ∈（第1个字段,...,第m个字段），得到第 j 条记录的第 i 个字段的数值 P，通过式(1)可以计算出来该值对应到第 i 个字段的第 nRange_{Bi} 个属性值区间内。将 nRange_{Bi} 的值，存入第 2*m+5+i 个字段(即 Tempi 字段)中。

(8)取出数组 N2 前 i-1 个数据进行叠加得到数值 n，从数组 N 中取出第 n 个数据来，存入第 m+5+i 个字段(即 Properi 字段)。

(9)如果 i<m，则返回第(7)步。

(10)如果 j<=表的记录数，则返回第(6)步。

(11)如果 j>表的记录数，则弹出“数据处理完毕”。

3 互信息提取

3.1 原理

D 分隔用来描述贝叶斯网络中的条件独立关系。对一个贝叶斯网络中任意 3 个互不连接的节点集合 X、Y 和 Z。如果在 X、Y 之间没有活动的连接路径，X、Y 被认为是通过 ZD 分隔的。其中，连接路径是两节点之间直接的连接弧段。

一个贝叶斯网络比那些仅涉及一个节点双亲的网络蕴含了更多的条件独立。如果对贝叶斯网络中的节点 $Node_i$ 和 $Node_j$ 之间的每个无向路径，在路径上有某个节点 $NodeIndex$，它有如下的 3 个属性之一，就说节点 $Node_i$ 和 $Node_j$ 条件独立于给定的节点集 E 即($I(Node_i, Node_j \mid E)$)。3 个属性是：

(1) $NodeIndex$ 在节点集 E 中，且路径上的两条弧都从 $NodeIndex$ 开始。

(2) $NodeIndex$ 在节点集 E 中，路径上的一条弧以 $NodeIndex$ 为头，另一个以 $NodeIndex$ 为尾。

(3) $NodeIndex$ 和它的任何后继都不在节点集 E 中，路径上的两条弧都以 $NodeIndex$ 为头。

给定节点集 E，当这些条件中的任何一个占据一条路径时，则称 $NodeIndex$ 阻塞那条路径。注意，在这个结果中引用的路径是无向路径，即路径忽略了弧方向。如果 $Node_i$ 和 $Node_j$ 之间的所有路径被阻塞，节点集 ED 分隔 $Node_i$ 和 $Node_j$ (依赖方向的分离)且得出结论：$Node_i$ 和 $Node_j$ 条件独立于给定的节点集 E。

D 分隔的概念也能应用到集合。给定集合 E，如果两个节点集 $Node_i$ 和 $Node_j$ 被 ED 分隔，则它们是条件独立的。给定 E，如果 $Node_i$ 中的所有节点和 $Node_j$ 中的所有节点之间的每条无向路径被阻塞，则 $Node_i$ 和 $Node_j$ 被 ED 分隔。即使使用了 D 分隔，在一般的贝叶斯网络中，概率推理是一个 NP 难题。然而，一般采用一个有向无环图(DAG)来简化贝叶斯网络。

根据信息论，两个离散随机变量(对应于节点)的 X 和 Y 具有联合概率函数 p (X , Y) 和边缘概率函数 p (X)、 p (Y)，其平均互信息 I (X , Y) 定义为

$$I(X, Y) = \sum_{X,Y} p(X, Y) \lg \frac{p(X, Y)}{p(X)\, p(Y)}$$

同样，条件互信息 I(X, Y|Z) 定义为

$$I(X, Y \mid Z) = \sum_{X,Y,Z} p(X, Y, Z) \lg \frac{p(X, Y \mid Z)}{p(X \mid Z)\, p(Y \mid Z)}$$

首先假设所有的节点之间存在连接，节点 X 和 Y 之间连接的潜在性运用条件互信息来计算。在通常情况下，设定一个较小正实数的阈值ε，当 $I(X, Y \mid Z) \le \varepsilon$ 时，称 X 与 Y 被条件集 Z 进行 d2 分割，即在给定 Z 的条件下，X 条件独立于 Y，从而删除 X 与 Y 之间的连接。经过 n(n-1)/2 次 CI 测试，最

后由完全潜在图修剪成稀疏的理想潜在图。

3.2 互信息计算算法步骤

功能 计算两个节点之间的互信息。

输入 节点编号 a，节点编号 b

输出 互信息值

算法：

For i 从 0 到 a 结点的 RangeWidth

{如果 a 不等于分类结点通过调用函数 CNormalMutmSingleProbablity(a，i)，计算 p(ai)的值；

否则，通过调用函数 CNormalMutmSingleProbablity2 (a，i)计算 p(ai) 的值；

如果 p(ai)的值小于等于 0

Continue；

For j 从 0 到 b 结点的 RangeWidth

{如果 b 不等于分类结点通过调用函数 CNormalMutmSingleProbablity(b，j)，计算 p(bj) 的值；

否则，通过调用函数 CNormalMutmSingleProbablity2 (b，j)，计算 p(bj) 的值；

如果 p(bj)的值小于等于 0

Continue；

如果 a 和 b 同时不为分类结点通过调用函数 CNormalMutmTwoProbablity(a，i，b，j)，计算 p(ai，bj)的值；

如果 a 为分类结点，而 b 不为分类结点，通过调用函数 CNormalMutmTwoProbablity2(b，j，i)，计算 p(ai，bj)的值；

如果 a 不为分类结点，而 b 为分类结点，通过调用函数 CNormalMutmTwoProbablity2(a，i，j)，计算 p(ai，bj)的值；

如果 p(ai，bj)的值小于等于 0

Continue

累计计算 p(ai，bj)*(log(p(ai，bj)/(p(ai)*p(bj)))) }

}

将累计计算的 p(ai，bj)*(log(p(ai，bj)/(p(ai)*p(bj))))的值返回

4 结论与讨论

本文阐述并实现的遥感图像数据分类预处理技术主要包括波谱空间分割和关系信息计算两部分。这种预处理技术既面向 BNs 模型构建，又具有一定的通用性，对其它智能数据分析技术同样具有比较强的借鉴意义。此外，数据预处理的时间复杂度还有待探讨，数据预处理对于模型构建的影响和效果笔者将另文阐述。

参考文献

1 李启青，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[J]. 电子信息学报, 2003, 10 (4): 1321-1327.

2 Park M, Stenstrom M K. Landuse Classification for Stormwater Modeling Using Baysian Networks[C]. Proc. of Diffuse Pollution Conference, Dublin, 2003.

3 曾祖麟. 粗集理论及其应用——关于数据推理的新方法[M]. 重庆: 重庆大学出版社, 1998.

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

5 Heckerman D. Bayesian Networks for Data Mining[J]. Data Mining and Knowledge Discovery, 1997, 1 (1): 79-119.

6 Nilsson N J. Artificial Intelligence: A New Synthesis[M]. Morgna Kaufmann Publishers Inc., 1998: 301-358.

~~~~~~

(上接第 260 页)

```
...
case KEY_ESC :
   keyflag=0;
   updown=0;
   if(menu==0)
   break;//如果已到主菜单，则不能再后退
do{
   for(k=0;k<menu;k++)
   { if(menuTable[k]==menu) break;}
   //本循环用于找父亲节点，找到即退出
   menu--;//准备找哥哥节点
   }while(k==menu+1);
//k==menu+1 代表未找到父亲，因为父亲节点的菜单值 menu 一
//定小于本节点(儿子节点)的 menu
menu=k;
MenuP=menu;
for(j=k;j>0;j--)
{
   for(n=j;n>=0;n--)
   if(menuTable[n]==k)
   break;//k 为父亲节点
   if(n>=0) break;// n>=0 则找到父亲结点
   k=k-1;
   //如果未找到，则准备找哥哥的父亲，直到找到大哥的父亲
//为止
}
//循环结束时 k 值即为要找的父亲节点的菜单值
updown=menu - k;
//父亲、大哥节点的菜单值之差即为菜单后退一级后新的光标的
//位置
DrowCaiDan(menu);//显示上级菜单
break;
```

4 菜单的建立和修改

使用本方法建立菜单时，应先画出菜单的树形拓扑结构，并给每个节点编号，此编号即为相应菜单选项的菜单值。编号时应由主菜单开始，从小到大依次编号，兄弟节点之间的编号应紧密相连，且“长辈”节点的菜单值应当小于“晚辈”节点的菜单值，它们之间可以有间隔，以备将来扩充菜单时使用。然后即可据此填写菜单数组。

修改菜单时，要遵循建立菜单时的规则，即先改菜单树，再改菜单表，最后调整 MenuDisplay( ) 函数即可。

5 结论

本文在分析常用的指针链表法实现树形拓扑结构方法的基础上，提出了用一维线性数组实现树形拓扑结构菜单的新方法，并用 C51 语言对该方法的可行性进行了验证。本方法对存储资源的利用效率高，通用性强，提高了系统的可操作性和应用的灵活性，当菜单显示内容改变时，只需修改相应参数即可达到目的。该方法适合在单片机应用系统中使用。

参考文献

1 黄声野，陈秀华，王东生. 一种用 C51 实现的单片机系统菜单管理方案[J]. 计算机工程, 2004, 30(3): 191-192.

2 彭良清. 基于节点编号的通用树状菜单设计方法与实现[J]. 单片机与嵌入式系统应用, 2002, 2(9): 36-40.

3 张培仁. 基于 C 语言编程 MCS－51 单片机原理与应用[M]. 北京: 清华大学出版社, 2003.

4 黄国瑜，叶乃青. 数据结构(C 语言版)[M]. 北京: 清华大学出版社, 2001: 116-121, 125-129.
~~~~~~